BAYESIAN MODELS ON MANIFOLDS FOR IMAGE REGISTRATION AND STATISTICAL SHAPE ANALYSIS

by

Miaomiao Zhang

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computing

School of Computing

The University of Utah

August 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of Miaomiao Zhang
has been approved by the following supervisory committee members:

Preston Thomas Fletcher , Chair	10/30/2015
	Date Approved
Guido Gerig , Member	11/10/2015
	Date Approved
Sarang Joshi , Member	10/30/2015
	Date Approved
Robert Michael Kirby , Member	10/30/2015
	Date Approved
Laurent Younes , Member	10/30/2015
	Date Approved

and by Ross T. Whitaker, Chair of the School of Computing
and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

An important area of medical imaging research is studying anatomical diffeomorphic shape changes and detecting their relationship to disease processes. For example, neurodegenerative disorders change the shape of the brain, thus identifying differences between the healthy control subjects and patients affected by these diseases can help with understanding the disease processes. Previous research proposed a variety of mathematical approaches for statistical analysis of geometrical brain structure in three-dimensional (3D) medical imaging, including atlas building, brain variability quantification, regression, etc. The critical component in these statistical models is that the geometrical structure is represented by transformations rather than the actual image data. Despite the fact that such statistical models effectively provide a way for analyzing shape variation, none of them have a truly probabilistic interpretation.

This dissertation contributes a novel Bayesian framework of statistical shape analysis for generic manifold data and its application to shape variability and brain magnetic resonance imaging (MRI). After we carefully define the distributions on manifolds, we then build Bayesian models for analyzing the intrinsic variability of manifold data, involving the mean point, principal modes, and parameter estimation. Because there is no closed-form solution for Bayesian inference of these models on manifolds, we develop a Markov Chain Monte Carlo method to sample the hidden variables from the distribution.

The main advantages of these Bayesian approaches are that they provide parameter estimation and automatic dimensionality reduction for analyzing generic manifold-valued data, such as diffeomorphisms. Modeling the mean point of a group of images in a Bayesian manner allows for learning the regularity parameter from data directly rather than having to set it manually, which eliminates the effort of cross validation for parameter selection. In population studies, our Bayesian model of principal modes analysis (1) automatically extracts a low-dimensional, second-order statistics of manifold data variability and (2) gives a better geometric data fit than nonprobabilistic models.

To make this Bayesian framework computationally more efficient for high-dimensional diffeomorphisms, this dissertation presents an algorithm, FLASH (finite-dimensional Lie

algebras for shooting), that hugely speeds up the diffeomorphic image registration. Instead of formulating diffeomorphisms in a continuous variational problem, Flash defines a completely new discrete reparameterization of diffeomorphisms in a low-dimensional bandlimited velocity space, which results in the Bayesian inference via sampling on the space of diffeomorphisms being more feasible in time. Our entire Bayesian framework in this dissertation is used for statistical analysis of shape data and brain MRIs. It has the potential to improve hypothesis testing, classification, and mixture models.

To my family

CONTENTS

LIST OF FIGURES

Figures

LIST OF TABLES

Tables

ACKNOWLEDGEMENTS

First of all, I would like to thank Prof. P. Thomas Fletcher for introducing me to the fields of image analysis and machine learning. His endless patience, continuous encouragement, and deep knowledge of science supported my entire Ph.D. study. *He is always the shining star in the sky, whose glory led me through the darkness and guided me to achieve my research goals.* I am so lucky to have had him as my adviser; he has cultivated the best in me as an independent researcher. I could not have imagined a better Ph.D. life than with him and his group. Secondly, I want to thank Prof. Laurent Younes for serving as my external committee member. His prominent work has inspired me to develop new ideas in medical image analysis. Moreover, special thanks to other committee members Prof. Guido Gerig, Prof. Sarang C. Joshi, and Prof. Robert M. Kirby for their valuable comments and insightful suggestions.

I would like to thank Dr. Nikhil Singh for contributing his time to discuss the fundamentals of differential geometry and diffeomorphic image registration. I also enjoyed many discussions of various kinds of research problems with my fellow lab mates - Dr. Wei Liu, Dr. Xiang Hao, Prasanna Muralidharan, Dr. Jacob Hinkle, and Dr. Bo Wang. Their different research backgrounds helped to broaden my horizons at the beginning of my Ph.D. program. I would like to thank Michelle Hromatka, Ting Liu, and Hang Shao, who worked with me. Their sense of humor filled our collaborations with joy. I also would like to thank Sam Preston, from whom I learned some GPU programming skills.

I would like to give special thanks to Prof. Ross Whitaker, who directed my first year of PhD study. I thank Prof. Martin Berzins for sharing his interesting hobbies and life experience with me, and Prof. Erin Parker for giving me valuable suggestions during my Ph.D. study. I also would like to thank my other lab mates Eleanor Wong, Gopal Veni, Yang Gao, and Dr. Anuja Sharma for helping me improve my presentations. I am deeply grateful to the Scientific Computing and Imaging Institute for providing a wonderful research atmosphere.

Finally, I am super thankful for my parents, Mr. Zhang and Ms. Wu, whose precious love surrounds me all the time. This dissertation would never had happened without their

trust and support. I will always love them. I thank all my other family members and friends who bring me laughter. Most of all, I feel so lucky to have found my love in the world - Mr. Bo Zhu. I have shared twelve years of happiness with him. We've been sustaining each other to realize our dreams all the way along. My life is full of sunshine and flowers because of him.

ACKNOWLEDGEMENTS IN CHINESE

首先感谢我的导师P. Thomas Fletcher教授带领我进入图像分析和机器学习领域. 他悉心的栽培, 源源不断的鼓励和渊博的科研知识给予了我整个博士学习最大的支持. 在我的研究生涯中, 他就像天空中的那颗明星, 指引我穿越黑暗到达目的地. 能成为他的学生, 我深感幸运. 他教会了我如何做最好的自己并成为一名独立的科研工作者. 没有他和他团队的带领, 我想我不会取得今天的成绩. 其次, 我要感谢约翰霍普金斯的Laurent Younes教授. 感谢他在百忙之中仍然成为我的委员会成员. 他在医学图像界杰出的贡献给予了我启发和灵感. 我更要感谢其他所有的委员会成员, Guido Gerig教授, Sarang C. Joshi教授和Robert M. Kirby教授. 感谢他们对我的工作提出的宝贵意见.

感谢Nikhil Singh博士和我一起探讨微分几何学和微分同胚图像配准问题. 同时我也非常开心能和其他的小组人员刘伟博士, 郝翔博士, Prasanna Muralidharan, Jacob Hinkle博士和王博博士讨论不同研究领域的问题. 他们扎实深厚的科研背景开阔了我博士学习初始阶段的视野. 在此, 我更要感谢我的合作伙伴Michelle Hromatka, 刘汀和邵航. 他们的聪明幽默风趣使我们的合作充满了欢乐. 我还要感谢Sam Preston, 从他那里我学到了GPU程序开发的一些基本知识.

特别感谢Ross Whitaker教授对我第一年的博士学习给予的指导. 感谢Martin Berzins教授常常跟我分享一些他有趣的爱好和人生阅历. 感谢Erin Parker教授在我博士学习和生活中给予的宝贵建议. 同时我还要感谢实验室的其他成员Eleanor Wong, Gopal Veni, 高阳和Anuja Sharma博士帮助我提高学术演说水平. 深深的感谢犹他大学的科学计算和图像所以及所有成员, 感谢他们营造了一个美好的学术氛围.

最后, 我要感谢我的父母张先生和吴女士. 他们对于我付出了所有的爱. 如果没有他们的相信和支持, 我不可能完成本篇博士论文. 我永远爱你们! 感谢我所有的亲人和朋友, 感谢他们给我的生活带来欢笑. 最后的最后, 我要感谢我是如此的幸运在这个诺大的世界中找到我唯一的爱 - 朱博先生. 十二年, 我们一起分享快乐. 我们一路彼此互相搀扶, 互相鼓励去实现自己的梦想. 我生活的每一天都充满了阳光和鲜花, 因为有你.

CHAPTER 1

INTRODUCTION

Statistical shape analysis develops methods for the geometric study of objects. The geometrical information is invariant to translation, scaling, and rotation. There is a vast range of applications of shape analysis in scientific research fields, including medical imaging, computer vision, machine learning, and biology. Defining a distance metric between shapes is an important first step of shape analysis. A distance metric allows one to estimate mean shapes and to extract shape variability from population-based data. Quantifying the shape variability of the human brain gives a direct interpretation of the anatomical brain structure that varies from one individual to another. In addition, brain shape changes relate to disease processes and changes in cognitive behavioral measures. For example, the shape and size of white and gray matter structures decrease in patients affected by neurodegenerative diseases, resulting in significantly larger corresponding brain ventricles. Shape analysis techniques have the potential to provide new insights and clinical assessments for disease diagnosis and treatment.

The means to represent shapes for a group of images is the geometric transformation between each individual and the mean image. Researchers can then perform statistical modeling of shape on the corresponding transformations. Finding a transformation that maps from one image to another, known as *deformable image registration*, is a basic area in the study of shape variation. In many applications, it is desirable that image transformations be diffeomorphisms, i.e., differentiable, bijective mappings with differentiable inverses. Such diffeomorphic mappings ensure several properties of the transformed images: (1) topology of objects in the image remains intact; (2) no nondifferentiable artifacts, such as creases or sharp corners, are created; and (3) the process can be inverted, for instance, to move back and forth between the source and target images. Computing the mean of images or atlas, which represents a large image dataset, is another important step for population-based shape analysis. The goal of atlas building is to compute a common coordinate system for anatomical comparisons across individuals. Other practical applications of atlas building

are alignment of functional data to a reference coordinate system and atlas-based image segmentation. A mean is only a point estimate and does not encode the variability of a population. Extracting low-dimensional, second-order statistics of data variability provides a compact representation of the anatomical variability in a large image database. Reducing the dimensionality to the inherent modes of shape variability has the potential to improve hypothesis testing, classification, and mixture models.

Previous methods to address these problems do not have truly probabilistic formulations. Having a probabilistic interpretation paves the way for factor analysis, mixture models, and generating stochastic systems. The main advantages of a probabilistic formulation are that (1) it considers uncertainty in the answers, for instance, 'a parameter has a probability of 0.95 of falling in a credible interval' and (2) it provides a natural and principled way to incorporate prior information into a model. Another important fact is that Bayesian analysis estimates model parameters directly from data rather than a hand-tuning manner, which eliminates the need for searching. Also, a Bayesian model is more robust when dealing with a small sample size of data or noisy samples, since the observations not following the general pattern are better modeled by the noise term.

This dissertation presents a novel Bayesian framework of statistical shape analysis and its application to diffeomorphic shape variability and brain MRI. We first carefully define manifold-valued likelihood distributions and prior distributions using the geodesic distance metric. We then develop Bayesian models for analyzing the intrinsic variability of manifold data and inference algorithms for estimating the mean point, model parameters, and principal modes. Due to the lack of closed-form solutions for Bayesian inference of these models, we develop Markov Chain Monte Carlo methods to sample the latent variables from the distribution and marginalize them over the posterior distribution. To make this Bayesian framework computationally efficient for high-dimensional diffeomorphisms, we propose a fast diffeomorphic image registration algorithm, FLASH.

While the practical applications introduced in this dissertation are mainly in the field of shape analysis on diffeomorphisms, the models and theory are generally applicable to generic manifold data, for example, vectors of unit length such as the sphere data, geometric transformations such as rotations and affine transforms, symmetric positive-definite tensors, and Stiefel manifolds (the set of orthonormal m-frames in Euclidean space $\mathbb{R}^n$). Although this dissertation focuses on applications to brain imaging, all statistical models are applicable to many other scientific research fields, such as machine learning, computer vision, and graphics. Several other possible applications will be discussed in the conclusion

and future work in Chapter 7.

1.1 Problems and Challenges

This section describes the problems and challenges of building these Bayesian models for statistical shape analysis on diffeomorphisms.

The first challenge for probabilistic modeling on diffeomorphisms is the lack of a closed-form solution for Bayesian inference. A previous method involving optimizing an objective function over the space of diffeomorphisms was a mode approximation of the posterior distribution, which performs poorly under image noise as shown in [2], even for a simple one-dimensional (1D) template estimation problem where the transformations are discrete shifts. Therefore, the diffeomorphisms should be treated as hidden random variables and not parameters to be estimated. In addition, despite the probabilistic motivation behind the practical problem, the model parameter that regularizes the smoothness of diffeomorphisms is not estimated in current practice, but rather specified in an ad hoc manner. This has several disadvantages: (1) it is difficult and time consuming to search for the appropriate parameters even for experienced modelers, (2) there is no certainty whether a good parameter is found or not, and (3) it is not clear whether the bad performance of a model with an assigned parameter is because no good parameter exists, or because it was not discovered.

The high dimensionality of the transformations, combined with the relatively small sample sizes available, makes statistical analysis difficult. Treating the dimensionality reduction of brain shape variability as a probabilistic inference problem on discrete images and jointly estimating the image atlas and principal geodesic modes of variation becomes the second challenge in shape variability analysis. Previous methods [3], [4] performed the dimensionality reduction after the fact, i.e., as a principal component analysis (PCA) of diffeomorphisms in the tangent space as a second stage after the estimation step. These models have two major disadvantages. First, they do not explicitly optimize the fit of the principal modes to the data intrinsically in the space of diffeomorphisms, resulting in a suboptimal fit to the data. Second, the number of dimensions to use is selected manually rather than inferred directly from the data. A related Bayesian model of PCA (BPCA) [5] for Euclidean data was proposed to automatically learn the dimension of the latent space from data by including a sparsity-inducing prior on each component of the factor matrix. This linear factor analysis model, however, is not applicable to nonlinear diffeomorphic transformations. The main difficulty of generalizing the model definition of BPCA to generic manifolds is the lack of an explicit formulation for the normalizing constant of distributions

on manifolds.

The last challenge of this Bayesian framework comes from the high computational complexity of its inference problem. Integrating out the high-dimensional diffeomorphisms by random sampling from the posterior distribution requires a huge number of computations on dense spatial grids that are prohibitively time consuming. One class of previous methods [1], [6] has been developed to reduce the large memory requirement, but the expensive computation of these methods still needs to be solved numerically on a full grid. Another class of diffeomorphic registration methods is the "greedy" algorithms [7], [8]. Greedy methods are much faster and more memory efficient, but they do not minimize a global variational problem and do not provide a distance metric between images. Thus, developing a fast diffeomorphic image registration algorithm without losing the distance metric of diffeomorphisms makes further statistical analysis computationally more feasible in time.

1.2 Dissertation Statement and Contributions

Thesis: *A **generative** Bayesian approach to analyze the data variability in nonlinear spaces provides parameter estimation and automatic dimensionality reduction for manifold data, such as diffeomorphisms. A Bayesian model of atlas building can for the first time estimate the parameter that regularizes the smoothness of diffeomorphisms. In population studies, (1) reparameterizing diffeomorphisms in a low-dimensional space with appropriate regularity parameters captures better intrinsic shape variability, and (2) having a discrete low-dimensional representation of diffeomorphisms makes model inference with Markov Chain Monte Carlo more feasible.*

To test this thesis statement, I have made the following contributions:

- **Bayesian estimation of regularization in diffeomorphic image registration.** We propose a truly probabilistic formulation of the diffeomorphic atlas building. This algorithm can for the first time estimate the regularity parameter of the smoothness of the diffeomorphisms simultaneously with other model parameters, including atlas and noise variance. We also discuss a mixture model of multiatlas estimation based on this setting.

- **Bayesian principal geodesic analysis of finite-dimensional manifolds.** We develop a latent variable model of principal geodesic analysis that provides a probabilistic framework for factor analysis on finite-dimensional manifolds. It shows that this model automatically selects the intrinsic dimensions of data variability, while giving a better fit to manifold data.

- **Bayesian principal geodesic analysis of infinite-dimensional diffeomorphisms.** We study a fully generative Bayesian formulation to jointly estimate model parameters and the low-dimensional latent space of diffeomorphisms in the population-based diffeomorphic image registration. The automatically selected latent dimensions from this model are able to reconstruct unobserved testing images with lower error than both linear principal component analysis (LPCA) in the image space and tangent space principal component analysis (TPCA) in the tangent space of diffeomorphisms.

- **Low-dimensional Lie algebras for fast diffeomorphic image registration.** To make the Bayesian inference of the models above computationally affordable, we define a novel discrete low-dimensional Lie algebra to approximate diffeomorphisms in a linear space. This not only speeds up a pairwise diffeomorphic image registration with a lot less demand on memory, but also accelerates the convergence rate to an optimal solution.

1.3 Outline

The remainder of this dissertation is organized as follows:

Chapter 2 introduces the background of basic Riemannian manifold concepts and diffeomorphic shape variability in medical image analysis. The first section briefly overviews Riemannian metrics, geodesics, Lie groups, and more; the second section describes fundamental tools for statistical shape analysis such as diffeomorphisms, diffeomorphic image registration, atlas building, and a Bayesian formulation of principal component analysis.

Chapter 3 presents a Bayesian framework of diffeomorphic atlas building to estimate the parameter that controls the diffeomorphic transformation regularity. A Monte Carlo Expectation Maximization algorithm (MCEM) is developed for inference where the expectation step is approximated via sampling on the manifold of diffeomorphisms. A mixture model of multiatlas building is then built and discussed in the context of this setting.

Chapter 4 describes a latent variable model for principal geodesic analysis that is analogous to probabilistic principal component analysis in Euclidean space. This model definition can be applied to any generic manifolds to discover low-dimensional factors using maximum likelihood.

Based on the setting presented in Chapters 3 and 4, Chapter 5 then presents a generative Bayesian approach for estimating the low-dimensional latent space of diffeomorphic shape variability in a population of images. A latent variable model is developed for principal geodesic analysis that provides a probabilistic framework for factor analysis in the space

of infinite-dimensional diffeomorphisms. A sparsity prior in the model results in automatic selection of the number of relevant dimensions by driving unnecessary principal geodesics to zero.

Chapter 6 introduces a fast geodesic shooting algorithm for a large deformation diffeomorphic metric mapping framework, named FLASH. This makes the Bayesian inference with Markov Chain Monte Carlo in Chapter 3 and Chapter 5 more feasible. A novel definition of low-dimensional Lie algebras in the space of bandlimited velocity fields is proposed, which results in most of the expensive computations needed for gradient descent methods entirely in these low-dimensional Lie algebras. FLASH not only speeds up the current diffeomorphic image registration algorithm dramatically, but also requires much less memory than state-of-the-art methods. FLASH is the first method to present a real discrete finite-dimensional Lie algebra structure to approximate infinite-dimensional diffeomorphisms. It breaks through the limitation of high computational cost and large memory demands of diffeomorphic atlas building for statistical shape variability analysis.

Chapter 7 concludes the dissertation with a general discussion and future work.

CHAPTER 2

BACKGROUND AND RELATED WORKS

2.1 Riemannian Manifolds

A Riemannian manifold is a smooth (differentiable) manifold equipped with a metric, which is a smoothly varying inner product on its tangent space. Riemannian manifolds arise naturally as the appropriate representations for data that have smooth constraints. For example, when analyzing directional data [9], i.e., vectors of unit length in $\mathbb{R}^n$, the correct representation is the sphere, S^{n-1}. Another important example of manifold data is in shape analysis, where the definition of the shape of an object should not depend on its position, orientation, or scale. Kendall [10] was the first to formulate a mathematically precise definition of shape as equivalence classes of all translations, rotations, and scalings of point sets. The result is a manifold representation of shape, or *shape space*. Linear operations violate the natural constraints of manifold data, e.g., a linear average of data on a sphere results in a vector that does not have unit length. As shown recently [11], using the kernel trick with a Gaussian kernel maps data onto a Hilbert sphere, and utilizing Riemannian distances on this sphere rather than Euclidean distances improves clustering and classification performance. Other examples of manifold data include geometric transformations, such as rotations and affine transforms, symmetric positive-definite tensors [12], [13], Grassmannian manifolds (the set of m-dimensional linear subspaces of $\mathbb{R}^n$), and Stiefel manifolds (the set of orthonormal m-frames in $\mathbb{R}^n$) [14].

This section gives a review of some necessary concepts about Riemannian manifolds that will be used later in the following chapters (see [15], [16] for more details).

2.1.1 Riemannian Metrics

Definition 1 *Let M be a smooth manifold, a **Riemannian metric** at each point $p \in M$ is an inner product $\langle \cdot, \cdot \rangle$ such that, for any two vector fields v, w on M, $\langle v, w \rangle_p \to R$. The norm of v is defined as $\|v\| = \langle v, v \rangle^{1/2}$.*

Under the definition of Riemannian metric, the length of a smooth curve $\gamma(t) : [0,1] \to M$ is

$$L(\gamma) = \int_0^1 \|\dot{\gamma}(t)\| dt, \tag{2.1}$$

where $\dot{\gamma}(t) = d\gamma(t)/dt$ is the tangent vector to $\gamma(t)$.

2.1.2 Geodesics

The shortest path between two points on a Riemannian manifold M is called a **geodesic**. It is a generalization of a straight line in Euclidean space to curved spaces. The geodesic flow is given by the variational problem [16], [17] of minimizing this energy functional

$$E(\gamma) = \int_0^1 \|\dot{\gamma}(t)\|^2 dt. \tag{2.2}$$

If γ is a geodesic, then

$$\frac{d}{dt}\langle \dot{\gamma}, \dot{\gamma} \rangle = 2\langle \nabla_{\dot{\gamma}} \dot{\gamma}, \dot{\gamma} \rangle = 0,$$

where $\nabla_{\dot{\gamma}} \dot{\gamma} = \frac{D}{dt}\dot{\gamma}$ is the covariant derivative of the vector field $\dot{\gamma}$.

Given two vector fields v, w, the **covariant derivative** $\nabla_v w$ generalizes the Euclidean directional derivative to a manifold setting and gives the change of the vector field w in the v direction. For example, consider a vector field $V(t)$ defined along γ, we can define the covariant derivative of V to be $\frac{DV}{dt} = \nabla_{\dot{\gamma}} V$. A vector field is called parallel if the covariant derivative along the curve γ is zero. A curve γ is geodesic if it satisfies the equation

$$\nabla_{\dot{\gamma}} \dot{\gamma} = 0.$$

In other words, geodesics are curves with zero acceleration.

For any point $p \in M$ and tangent vector $v \in T_p M$, the tangent space of M at p, there is a unique geodesic curve γ with initial conditions $\gamma(0) = p$ and $\dot{\gamma}(0) = v$. This geodesic is guaranteed to exist only locally. A **Riemannian exponential map** takes the position p and velocity v as input and returns the point at time 1 along the geodesic with these initial conditions (see Figure 2.1). When γ is defined over the interval $[0,1]$, the Riemannian exponential map at p is defined as

$$\mathrm{Exp}(p, v) = \gamma(1).$$

The exponential map is locally diffeomorphic onto a neighbourhood of p. Let $V(p)$ be the largest such neighbourhood. The inverse of the exponential map exists within $V(p)$. This

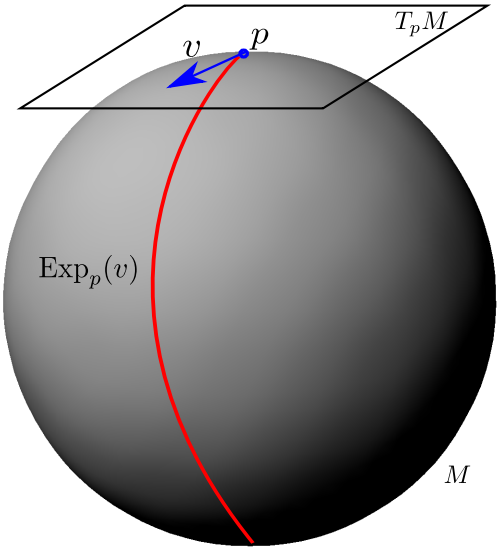


Figure 2.1: Riemannian exponential map.

map is called **Riemannian log map** $\mathrm{Log}_p : V(p) \to T_pM$. For any point $q \in V(p)$, the Riemannian distance function is given by

$$d(p, q) = \|\mathrm{Log}(p, q)\|.$$

For the purpose of notation simplicity, the point p will be included as a parameter in the exponential and log maps, i.e., define $\mathrm{Exp}(p, v) = \mathrm{Exp}_p(v)$ and $\mathrm{Log}(p, q) = \mathrm{Log}_p(q)$.

Example 1 *Let p be a point on an n-dimensional sphere embedded in $\mathbb{R}^{n+1}$, and let v be a tangent at p. The inner product between tangents at a base point p is the usual Euclidean inner product. The exponential map is given by a 2D rotation of p by an angle given by the norm of the tangent, i.e.,*

$$\mathrm{Exp}(p, v) = \cos\theta \cdot p + \frac{\sin\theta}{\theta} \cdot v, \quad \theta = \|v\|.$$

The log map between two points p, q on the sphere can be computed by finding the initial velocity of the rotation between the two points. Let $\pi_p(q) = p \cdot \langle p, q \rangle$ denote the projection of the vector q onto p. Then,

$$\mathrm{Log}(p, q) = \frac{\theta \cdot (q - \pi_p(q))}{\|q - \pi_p(q)\|}, \quad \theta = \arccos(\langle p, q \rangle).$$

2.1.3 Lie Groups

A Lie group is both a smooth manifold and group in which all the group operations are smooth mappings. The theory of Lie groups plays a vital role in medical image registration,

population shape analysis, etc. This section studies the basic properties of Lie groups. More details can be found in [15], [18], [19].

Definition 2 *A **group** G is a set of elements with a binary operation $\cdot$, such that*

- $\forall x, y \in G$, $x \cdot y \in G$.
- $\forall x, y \in G$, $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.
- *There exists a unique element $e \in G$, such that $\forall x \in G$, $e \cdot x = x \cdot e = x$ holds.*
- *For each $x \in G$, there is an element $y \in G$ such that $x \cdot y = y \cdot x = e$.*

Definition 3 *A **Lie group** is a smooth manifold equipped with group structures, in which the group multiplication and inversion*

$$(x, y) \longmapsto x \cdot y \ : \ G \times G \to G$$
$$x \longmapsto x^{-1} \ : \ G \to G$$

are both smooth mappings.

Example 2 *A very simple example is the real numbers $\{\mathbb{R}\}$ with the group operation addition. The element identity is 0, and the inverse of each element $x \in \mathbb{R}$ is $-x$.*

Example 3 *A **matrix group** is a Lie subgroup of a general linear group $GL(n; \mathbb{R})$ over real numbers. It is a group of all $n \times n$ invertible matrices under matrix multiplication. The identity element is the $n \times n$ identity matrix, and the inverse of an element is matrix inverse.*

Every Lie group is associated with a corresponding space of infinitesimal transformations, known as Lie algebra, which is the tangent space of the Lie group at identity. It represents a local structure of the Lie group.

Definition 4 *A **Lie algebra** is a vector space $\mathfrak{g}$ together with a Lie bracket $[\cdot, \cdot]$ that maps $\mathfrak{g} \times \mathfrak{g}$ to $\mathfrak{g}$. $\forall x, y, z \in \mathfrak{g}$ and $a, b \in \mathbb{R}$, the following axioms are satisfied:*

(a) Linearity: $[ax + by, z] = a[x, z] + b[y, z]$,

(b) Anticommutativity: $[x, y] = -[y, x]$,

(c) Jacobi identity: $[x, [y, z]] + [z, [x, y]] + [y, [z, x]] = 0$.

The Lie bracket is a nonassociative multiplication operator. Two vectors x and y of a Lie algebra commute if $[x, y] = 0$.

2.1.4 Left and Right Invariant Metrics

Given a smooth mapping $\varphi : M \to N$ between manifolds M and N with Riemannian metric $\langle \cdot, \cdot \rangle$, a **pullback metric** $\varphi^* \langle \cdot, \cdot \rangle$ on M is induced by φ as

$$\varphi^* \langle v, w \rangle_p = \langle \varphi(v_p), \varphi(w_p) \rangle.$$

Definition 5 *A Riemannian metric $\langle \cdot, \cdot \rangle$ on Lie group G is called **left-invariant** if it is invariant under left multiplication such as*

$$\langle \cdot, \cdot \rangle = (L_g)_* \langle \cdot, \cdot \rangle,$$

for all $g \in G$.

*Similarly, a Riemannian metric is **right-invariant** if it is invariant under right multiplication, which is $\langle \cdot, \cdot \rangle = (R_g)_* \langle \cdot, \cdot \rangle$.*

2.1.5 Jacobi Fields

Consider a variation of geodesics $\gamma(s, t) : (-\epsilon, \epsilon) \times [0, 1] \to M$, with initial conditions $\gamma(0, t) = p$ and $\gamma(p, 0) = v$. The variation $\gamma(s, t) = \mathrm{Exp}(p, su + tv)$ where $u \in T_p M$, produces a "fan" of geodesics. This is illustrated for a sphere in Figure 2.2. Taking the derivative of

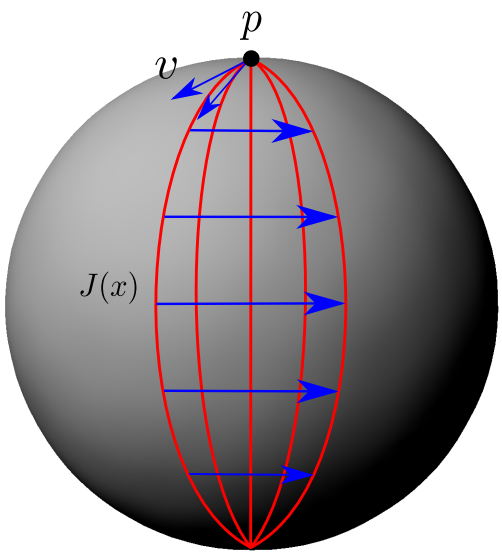


Figure 2.2: Jacobi fields

this variation results in a Jacobi field: $J_v(t) = d\gamma/ds(0,t)$. Finally, an expression for the exponential map derivative is given as

$$d_v\mathrm{Exp}(p,v) = J_v(1). \tag{2.3}$$

For a general manifold, computing the Jacobi field $J(t)$ requires solving a second-order ordinary differential equation as

$$\nabla^2_{\dot\gamma(t)} J(t) + R(\dot\gamma(t), J(t))\dot\gamma(t) = 0, \tag{2.4}$$

where R denotes a Riemannian curvature tensor that expresses the curvature of a Riemannian manifold M.

There is a reduced version of Jacobi fields under invariant Riemmanian connections from Bullo [20], which is also used by Hinkle et al. [21]. The second-order equations are reduced to a set of first-order equations, which are straightforward to linearize and more easily find a solution of the original system through the simplified system.

Under the left invariant metric of Lie groups, we define a vector field $U(t)$ as a left trivialized Jacobi field $U(t) = \gamma^{-1}J(t)$, and a variation of the left trivialized velocity v along the geodesic is δv. These variables satisfy the following simple reduced Jacobi equations:

$$\frac{d}{dt}\begin{pmatrix} U \\ \delta v \end{pmatrix} = \begin{pmatrix} -\mathrm{ad}_v & I \\ 0 & -\mathrm{sym}_v \end{pmatrix}\begin{pmatrix} U \\ \delta v \end{pmatrix}, \tag{2.5}$$

where $\mathrm{sym}_v\delta v = -\mathrm{ad}^\dagger_v\delta v - \mathrm{ad}^\dagger_{\delta v}v$, and $\mathrm{ad}^\dagger$ denotes the adjoint operator of ad.

The reduced adjoint Jacobi fields are then simply computed by the adjoint of the reduced Jacobi equations in (2.5). This results in another ordinary differential equation (ODE) as

$$\frac{d}{dt}\begin{pmatrix} \hat U \\ \delta\hat v \end{pmatrix} = \begin{pmatrix} \mathrm{ad}^\dagger_v & 0 \\ -I & \mathrm{sym}^\dagger_v \end{pmatrix}\begin{pmatrix} \hat U \\ \delta\hat v \end{pmatrix}, \tag{2.6}$$

where $\hat U, \delta\hat v \in V$ are introduced adjoint variables, and $\mathrm{sym}^\dagger_v\delta\hat v = -\mathrm{ad}_v\delta\hat v + \mathrm{ad}^\dagger_{\delta\hat v}v$. For more details on the derivation of the reduced adjoint Jacobi field equations, see [20].

Similarly, the reduced Jacobi equations and adjoint Jacobi equations under a right invariant metric are

$$\frac{d}{dt}\begin{pmatrix} U \\ \delta v \end{pmatrix} = \begin{pmatrix} \mathrm{ad}_v & I \\ 0 & \mathrm{sym}_v \end{pmatrix}\begin{pmatrix} U \\ \delta v \end{pmatrix}, \tag{2.7}$$

$$\frac{d}{dt}\begin{pmatrix} \hat U \\ \delta\hat v \end{pmatrix} = \begin{pmatrix} -\mathrm{ad}^\dagger_v & 0 \\ -I & -\mathrm{sym}^\dagger_v \end{pmatrix}\begin{pmatrix} \hat U \\ \delta\hat v \end{pmatrix}. \tag{2.8}$$

The Jacobi fields can be evaluated in closed form for the class of manifolds known as *symmetric spaces*, see for instance [16]. The explicit formulas of Jacobi field computations

exist for Riemannian symmetric spaces, such as compact Lie groups, sphere, Kendall shape spaces, and Grassmannians. Before introducing the definition of symmetric space, we first review an isometry of a Riemannian manifold M.

Definition 6 *Let a mapping $f : M \to N$ be a diffeomorphism between Riemannian manifolds M and N. An **isometry** is a distance-preserving function if for any $x \in M$ it has*

1. *The derivative of f at x is an isomorphism of tangent space $df_x : T_xM \to T_{f_x}M$.*
2. *For any $v, w \in T_xM$, the Riemannian metric preserves as $\langle v.w \rangle_M = \langle df_x(v), df_x(w) \rangle_N$.*

Definition 7 *A Riemannian manifold M is a **symmetric space** if $\forall x \in M$, there exists an isometry of M, s_x, and a neighborhood N_x of x where x is the unique fixed point of s_x in N_x.*

Example 4 *A very simple example is Euclidean space $\mathbb{R}^n$ with the Euclidean metric. For any point $x \in \mathbb{R}^n$, there is $s_x(x + v) = x - v$. The isometry group is the Euclidean group $E(n)$ generated by translations.*

2.1.6 Adjoint Representation

Given any element $g \in G$, a conjugation mapping $\Psi : G \to G$ is defined as

$$\Psi_g(h) = ghg^{-1},$$

for all $h \in G$.

Definition 8 *The **adjoint representation of a Lie group** is the derivative of $\Psi_g(h)$ with respect to h at the identity, which is*

$$d(\Psi_g)_e = \mathrm{Ad}_g : G \times \mathfrak{g} \to \mathfrak{g}.$$

Taking the derivative of the adjoint map Ad *w. r. t. g at the identity gives the **adjoint representation of a Lie algebra** $\mathfrak{g}$:*

$$d(\mathrm{Ad}_g)_e = \mathrm{ad} : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}.$$

Example 5 *For any $g, h \in GL(n;\mathbb{R})$, the Ad_g is derived as*

$$\begin{aligned}\mathrm{Ad}_g w &= \frac{\partial}{\partial \xi}\Psi_g(h_\xi)\big|_{\xi=0},\\ &= g\frac{\partial}{\partial \xi}h_\xi\big|_{\xi=0}g^{-1},\\ &= gwg^{-1},\end{aligned}$$

where h_ξ denotes the variation of h by ξ such that $h_0 = e$ and $\frac{\partial}{\partial \xi}h_\xi\big|_{\xi=0} = w$ for $\Psi_g(h_\xi) = gh_\xi g^{-1}$.

Similarly, we then take the derivative w. r. t. g and arrive at

$$\begin{aligned}\mathrm{ad}_v w &= \frac{\partial}{\partial \xi}\mathrm{Ad}_{g_\xi} w\Big|_{\xi=0},\\ &= \frac{\partial}{\partial \xi}(g_\xi w g_\xi^{-1})\big|_{\xi=0},\\ &= \left(\frac{\partial}{\partial \xi}g_\xi w g_\xi^{-1}\right)\Big|_{\xi=0} + \left(g_\xi w \frac{\partial}{\partial \xi}g_\xi^{-1}\right)\Big|_{\xi=0},\\ &= vw - \left(g_\xi w g_\xi^{-1}\frac{\partial}{\partial \xi}g_\xi g_\xi^{-1}\right)\Big|_{\xi=0},\\ &= vw - wv,\end{aligned}$$

where g_ξ is the variation of g by ξ with $g_0 = e$ and $\frac{\partial}{\partial \xi}g_\xi\big|_{\xi=0} = v$ for $\mathrm{Ad}_{g_\xi} w = g_\xi w g_\xi^{-1}$.

2.2 Statistics on Manifolds

This section introduces basic statistical models on manifolds that we will use in the following chapters.

2.2.1 Bayesian Principal Component Analysis

Principal component analysis (PCA) [22] has been widely used to analyze the high-dimensional data. A common way of describing PCA is to find a subspace through the data mean that represents the maximal total variance of the original dataset. Consider a set y of n-dimensional Euclidean random variables $\{y_j\}_{j=1,\dots,N} \in \mathbb{R}^n$. The mean is computed by a linear average over the data as

$$\mu = \frac{1}{N}\sum_{j=1}^{N} y_j.$$

The sample covariance matrix S is then computed as

$$S = \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \mu)(y_j - \mu)^T.$$

To develop a latent variable model for PCA that provides a probabilistic framework for factor analysis, Tipping and Bishop proposed probabilistic PCA (PPCA) [23]. A similar

formulation was proposed by Roweis [24]. Other examples of latent variable models include probabilistic canonical correlation analysis (CCA) [25] and Gaussian process models [26].

The main idea of PPCA is to model a set y of n-dimensional Euclidean random variables $\{y_j\}_{j=1,\ldots,N} \in \mathbb{R}^n$. The relationship between each variable y_j and its corresponding q-dimensional ($q < n$) latent variable x_j is

$$y_j = \mu + Bx_j + \epsilon, \tag{2.9}$$

where μ is the mean of dataset $\{y_j\}$, x_j is conventionally defined as a random variable generated from $N(0, I)$, B is an $n \times q$ factor matrix that relates x_j and y_j, and $\epsilon \sim N(0, \sigma^2 I)$ represents error. This definition gives a data likelihood as

$$p(y \mid x; B, \mu, \sigma) \propto \prod_{j=1}^{N} \exp\left(-\frac{\|y_j - \mu - Bx_j\|^2}{2\sigma^2}\right).$$

PPCA opened up the possibility for probabilistic interpretations for different kinds of factor analyses. Bishop [5] extended PPCA by adding a prior on the factors, resulting in automatic selection of model dimensionality, called Bayesian PCA (BPCA). Figure 2.3 shows the graphical model of BPCA. It automatically learns the dimension of the latent space from data by including a Gaussian prior over each column of the factor matrix B, which is known as an automatic relevance determination (ARD) prior. Each such Gaussian has an independent variance associated with a precision hyperparameter γ_i, so that

$$p(B \mid \gamma) = \prod_{i=1}^{q} \left(\frac{\gamma_i}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\gamma_i B_i^T B_i\right),$$

where B_i denotes the ith column of B.

The value of γ_i is estimated iteratively as $\frac{n}{\|B_i\|^2}$ in this model, and thus enforces sparsity by driving the corresponding component B_i to zero. More specifically, if γ_i is large, B_i will be effectively removed in the latent space. This arises naturally because the larger γ_i is, the lower the probability of B_i will be. Notice that the columns of B define the principal subspace of standard PCA, therefore, inducing sparsity on B has the same effect as removing irrelevant dimensions in the principal subspace.

2.2.2 Principal Geodesic Analysis

The previous section describes the linear statistical models to analyze the underlying dimensionality of data variability. None of these approaches are applicable to the manifold data in nonlinear spaces. *Principal geodesic analysis* proposed by Fletcher et al. [27] generalizes the commonly used technique in linear spaces, principal component analysis,

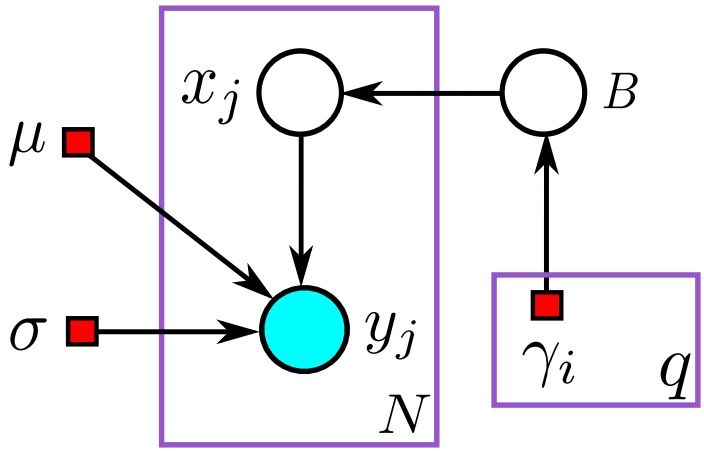


Figure 2.3: Graphical model of BPCA.

to generic manifolds. It describes the geometric variability of manifold data by finding lower-dimensional geodesic subspaces that minimize the residual sum-of-squared geodesic distances to the data.

2.2.2.1 Means on Manifolds

Analogous to principal component analysis in Euclidean space, the first important step for principal geodesic analysis is to compute the Fréchet mean [28], which is a general definition of the mean point on manifold M. More details of the Fréchet mean can be found in [29], [30]. Given a collection of points $\{y_j\}_{j=1,\dots,N} \in M$, the Fréchet mean of this dataset is formulated in a minimization problem as

$$\arg\min_{\mu} \frac{1}{2N} \sum_{j=1}^{N} d(\mu, y_j)^2. \tag{2.10}$$

Notice that this energy function depends on the definition of Riemannian distance $d(\cdot,\cdot)$ on the manifold M. Thus, a different distance metric on a different manifold obtains a different formulation for the mean point.

Following [27], [29], [31], we use a gradient descent algorithm to optimize the equation 2.10 with its gradient w. r. t. the mean point μ as

$$-\frac{1}{N} \sum_{j=1}^{N} \mathrm{Log}_{\mu}(y_j).$$

With the current estimate $\hat{\mu}$, the equation for updating the mean is

$$\mu^{\mathrm{new}} = \mathrm{Exp}_{\hat{\mu}} \left(\frac{\tau}{N} \sum_{j=1}^{N} \mathrm{Log}_{\hat{\mu}}(y_j) \right),$$

where τ is the step size.

2.2.2.2 Variance on Manifolds

We use the definition of variance from Fréchet [28] for general manifolds. For a random variable y in a metric space with Fréchet mean μ, the variance is given by the expected value of the squared distance to the mean point as

$$\sigma^2 = \mathbb{E}[d(\mu, y)^2].$$

Therefore, the sample variance of a dataset $\{y_j\}_{j=1,\dots,N}$ is

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N} d(\mu, y_j)^2 = \frac{1}{N}\sum_{j=1}^{N} \|\mathrm{Log}_\mu(y_j)\|^2.$$

This definition coincides with the standard Euclidean variance when $M = \mathbb{R}^n$.

Fletcher et al. [32] then proved the principal directions and variances can be simply computed from the covariance matrix

$$S = \frac{1}{N-1}\sum_{j=1}^{N} \mathrm{Log}_\mu(y_j)\mathrm{Log}_\mu(y_j)^T.$$

In summary, the algorithm for principal geodesic analysis on manifold data is shown in Algorithm 1.

Algorithm 1: Principal Geodesic Analysis

Input: Dataset $\{y_j\}_{j=1,\dots,N} \in M$.
μ: Fréchet mean of dataset, $v_q \in T_\mu M$: principal directions, $\lambda_q \in \mathbb{R}$: variances.
$u_j = \mathrm{Log}_\mu(y_j)$,
$S = \frac{1}{N-1}\sum_{j=1}^{N} u_j u_j^T$,
Output: $\{v_q, \lambda_q\}$ as eigenvectors, eigenvalues of S.

2.2.3 Geodesic Regression and Splines

The goal of *geodesic regression* [33], [34] is to find the relationship between an independent variable $x \in \mathbb{R}$ and a manifold-valued dependent random variable $y \in M$. Unlike the linear regression in Euclidean space, this relationship of geodesic regression is modeled as a geodesic curve on the manifold. In other words, it directly generalizes the linear regression to the manifold. Niethammer et al. [35] independently proposed geodesic regression for the case of diffeomorphic transformations of image time series. Hinkle et al. [36] used constant higher-order covariant derivatives to define intrinsic polynomial curves on a Riemannian

manifold for regression. Nonparametric kernel regression on Riemannian manifolds has also been proposed by Davis et al. [37]. Shi et al. [38] proposed a semiparametric model for manifold response data, which also has the ability to handle multiple covariates.

For each given data point $y_j \in M$ associated with $x_j \in \mathbb{R}$, we model it in the form of geodesic regression as

$$y_j = \mathrm{Exp}(\mathrm{Exp}(\mu, wx_j), \epsilon), \tag{2.11}$$

where μ is the mean point, $w \in T_\mu M$ is a tangent vector at μ, and ϵ is the noise variance.

To estimate the parameters $\{\mu, w\}$, we formulate geodesic regression as a least square problem. A sum-of-squared error of the manifold data from the geodesic is given by

$$E(\mu, w) = \frac{1}{2}\sum_{j=1}^{N} d(\mathrm{Exp}(\mu, wx_j), y_j)^2.$$

Again, note that this energy function coincides with the ordinary least square problem in Euclidean space when $M = \mathbb{R}^n$. However, this optimization problem does not typically yield an analytic solution due to the fact that the derivative of $\mathrm{Exp}(\cdot, \cdot)$ is computationally complicated in nonlinear spaces. Fletcher et al. [32] developed a gradient descent algorithm to search for the optimal answer.

A problem closely related to the regression problem is that of fitting smoothing splines to manifold data. The typical objective function for smoothing splines is a combination of a data matching term and a regularization term for the spline curve. For example, Su et al. [39] proposed a smoothing spline where the data matching is the same least squares objective as the regression problem (2.11), leading to a smoothing splines optimization. Jupp and Kent [40] proposed solving the smoothing spline problem on a sphere by unrolling onto the tangent space. This unrolling method was later extended to shape spaces by Kume [41]. Smoothing splines on the group of diffeomorphisms has been proposed as growth models by Miller et al. [42] and as second-order splines by Trouvé et al. [43]. A similar paradigm is used by Durrleman et al. [44] to construct spatiotemporal image atlases from longitudinal data. Yet another related problem is the spline *interpolation* problem, where the data matching term is dropped and the regularization term is optimized subject to constraints that the curve pass through specific points. The pioneering work of Noakes et al. [45] introduced the concept of a cubic spline on a Riemannian manifold for interpolation. Crouch and Leite [46] investigated further variational problems for these cubic splines and for specific classes of manifolds, such as Lie groups and symmetric spaces. Buss and Fillmore [47] defined interpolating splines on the sphere via weighted Fréchet averaging.

2.3 Diffeomorphic Image Registration

A **diffeomorphism** is a bijective smooth mapping with a smooth inverse. The space of diffeomorphisms is an infinite-dimensional Lie group that has been studied for statistical shape analysis of medical images. Such a diffeomorphic mapping prevents folding or tearing of the grid that might occur in small deformation image registration methods. The goal of the deformable template approach to statistical shape analysis of images is to quantify shape using deformable image registration and then compute the statistics of the resulting diffeomorphisms, rather than the images themselves. Computing a template image, or atlas, which represents a large dataset is the first step in this process. The class of diffeomorphic mapping functions first brings group images in different coordinate systems into a common space, such that anatomical comparisons and population-based analysis can be performed across individuals. Also, the diffeomorphism preserves the topology of objects in the images and provides forward and inverse mappings between the atlas and individuals.

This section gives a brief background of diffeomorphisms, diffeomorphic image registration, diffeomorphic atlas building, and a Bayesian formulation of principal component analysis for Euclidean data that automatically reduces the dimensionality of a high-dimensional dataset.

2.3.1 Diffeomorphisms

Consider images to be square-integrable functions defined on a d-dimensional torus domain $\Omega = \mathbb{R}^d/\mathbb{Z}^d$, that is, an image is an element of $L^2(\Omega, \mathbb{R})$. A diffeomorphism is a bijective invertible mapping $\phi : \Omega \to \Omega$ with its smooth inverse ϕ^{-1}. We denote the space of diffeomorphisms as $\mathrm{Diff}(\Omega)$ whose derivatives infinitely exist and are square-integrable, as is the inverse mapping.

The *Lie algebra* of the diffeomorphism group consists of all smooth vector fields on Ω equipped with the *Lie bracket* of vector fields. In other words, we can define a Lie algebra on the tangent space of diffeomorphisms $V = T\mathrm{Diff}(\Omega)$ with a known Lie bracket operation. Most of the computations with respect to Lie groups are done in Lie algebra since it is a linear space that has nice properties to work with.

Given a flow of time-varying velocity field, $v_t : [0, 1] \to V$, we generate diffeomorphisms $t \mapsto \phi_t \in \mathrm{Diff}(\Omega)$ (see Figure 2.4) as a solution to the following ordinary differential equation

$$\frac{d\phi_t}{dt}(x) = v_t(x) \circ \phi_t(x). \tag{2.12}$$

Note that we use subscripts for the time variable, i.e., $v_t(x) = v(t, x)$, and $\phi_t(x) = \phi(t, x)$.

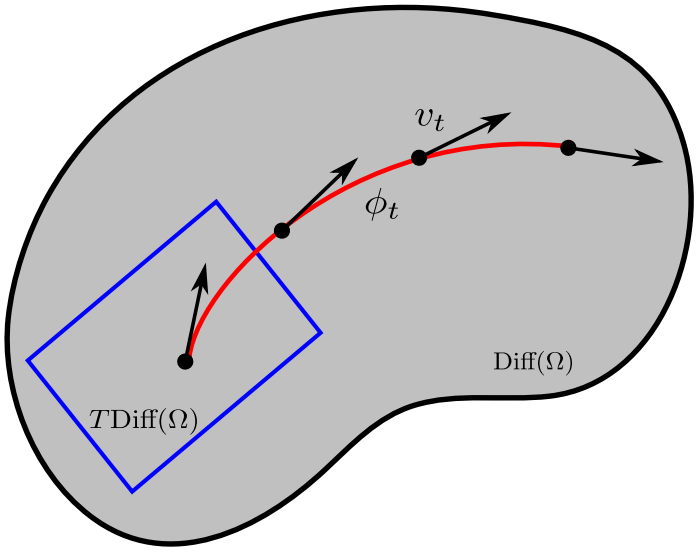


Figure 2.4: Space of diffeomorphisms.

2.3.2 Metrics on Diffeomorphisms

A key ingredient in computational anatomy is the notion of a distance metric on the space of diffeomorphisms. Such a metric provides a means for quantifying the magnitude of the deformation between two images, and forms the mathematical foundation for estimation of statistical models, such as atlases, as least-squares minimization problems. The first step is to define an inner product on the space of velocities, $V = T_e\text{Diff}(\Omega)$, identified with the tangent space at the identity transform, $e \in \text{Diff}(\Omega)$. This inner product is of the form

$$\langle v, w \rangle_V = \int_\Omega \langle Lv(x), w(x) \rangle dx,$$

for $v, w \in V$, and a symmetric, positive-definite differential operator $L : V \to V^*$, mapping to the dual space, V^*. We use $L = (-\alpha\Delta + I)^c$, for some constant $\alpha > 0$ and integer power c. The dual to the vector v is a momentum, $m \in V^*$, such that $m = Lv$ or $v = Km$, where K is the inverse of L.

Next we define a right-invariant metric as an inner product at any other point $\phi \in \text{Diff}(\Omega)$ by pulling back the velocities at ϕ to the identity by right composition. In other words, for $v_t, w_t \in T_\phi\text{Diff}(\Omega)$ the right-invariant metric is given by

$$\langle v_t, w_t \rangle_{T_\phi\text{Diff}(\Omega)} = \langle v_t \circ \phi^{-1}, w_t \circ \phi^{-1} \rangle_V.$$

A geodesic curve $\{\phi_t\} \in \text{Diff}(\Omega)$, illustrated in Figure 2.5, is computed through an energy

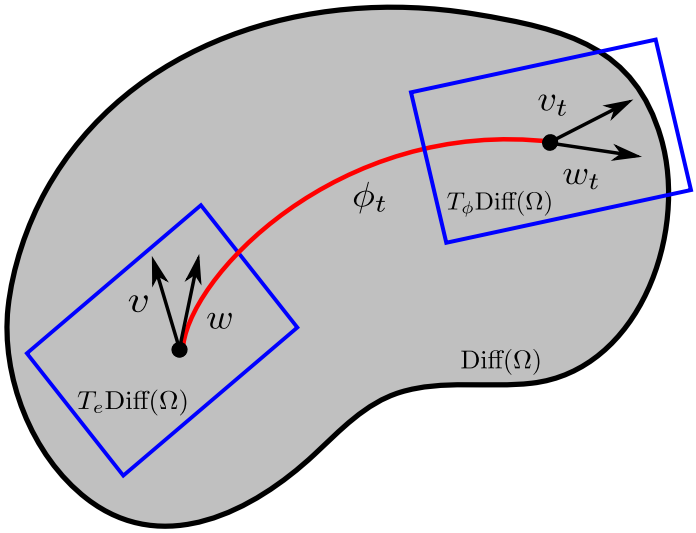


Figure 2.5: Metrics on diffeomorphisms.

minimization problem as

$$E(\phi_t) = \int_0^1 \left\| \frac{d\phi_t}{dt} \circ \phi_t^{-1} \right\|_V^2 dt,$$

and characterized by the Euler-Poincaré equations (EPDiff) [48], [49],

$$\begin{aligned} \frac{\partial v}{\partial t} &= -\mathrm{ad}_v^\dagger v = -K \mathrm{ad}_v^* m \\ &= -K \left[(Dv)^T m + Dm\, v + m \,\mathrm{div}\, v \right], \end{aligned} \tag{2.13}$$

where D denotes the Jacobian matrix. The operator ad^* is the dual of the negative Lie bracket of vector fields,

$$\mathrm{ad}_v w = -[v, w] = Dvw - Dwv. \tag{2.14}$$

Given an initial velocity, $v_0 \in V$, at $t = 0$, the EPDiff equation (2.13) can be integrated forward in time, resulting in a time-varying velocity $v_t : [0, 1] \to V$, which itself is subsequently integrated in time by the rule $(d\phi_t/dt) = v_t \circ \phi_t$ to arrive at the geodesic path, $\phi_t \in \mathrm{Diff}(\Omega)$. This process is known as *geodesic shooting*.

2.3.3 LDDMM With Geodesic Shooting

Several works [50], [51] modeled the flow of diffeomorphism by integrating over its time-dependent velocity field, using Lagrange transport equations. Later, Beg et al. [52] proposed an elegant mathematical formulation, known as *large deformation diffeomorphic metric mapping* (LDDMM), showing that the velocity field over time generates diffeomorphisms for large deformation diffeomorphic image registration. This framework introduced a distance

metric on the space of diffeomorphisms between images, which gave rise to a variational principle that expresses the optimal image registration as a geodesic flow. The advantages of having a distance metric are: (1) it formulates a statistical model of the least square problem via minimization of the sum-of-squared residual distance, for example, the model proposed by Zhang et al. [53]; (2) because this distance between images encodes the information of geometric variability, a number of theoretical methods related to LDDMM, especially the ones in the statistical analysis of anatomical shapes (for instance, longitudinal analysis, group comparisons, etc.), were further developed in [3], [54], [55].

In LDDMM, an image I_0 is deformed by a diffeomrophism ϕ as $I_0 \circ \phi^{-1}$. Figure 2.6 depicts an example of transforming a brain image.

Given a source image I_0 and a target image I_1, we minimize an energy function of sum-of-squared distance function plus regularization term to estimate the diffeomorphic transformation as

$$E(v_t) = \frac{1}{2\sigma^2}\|I_0 \circ \phi_1^{-1} - I_1\|_{L^2}^2 + \int_0^1 \|v_t\|_V^2 \, dt, \tag{2.15}$$

where σ^2 represents image noise variance.

A variational scheme was described in [52] to simulate the evolution of velocity v_t and diffeomorphism ϕ_t at each discrete time point using gradient descent. In order to store the entire flow of time-varying velocities and diffeomorphisms, this approach requires a large amount of memory. Vialard et al. [1] proposed to estimate only the initial velocity v_0 instead of the whole sequence of v_t by geodesic shooting. They also derived a backward integration of adjoint equations to carry gradients of the image matching term at time

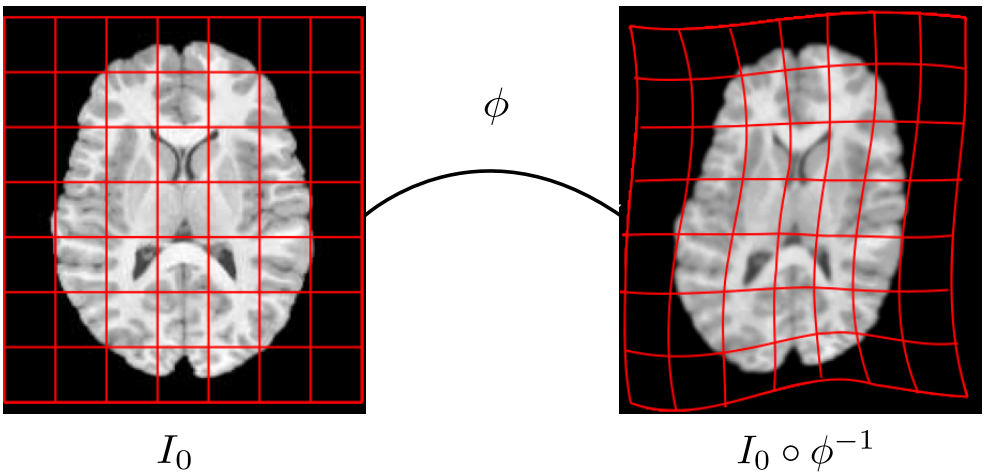


Figure 2.6: Deforming an axial of a 3D brain MRI image by ϕ.

point $t = 1$ back to the initial velocity at $t = 0$. Details can be found in [1], [56]. Since the geodesic is uniquely determined by the initial velocity v_0, we rewrite the LDDMM energy function (2.15) with respect to the initial velocity as

$$E(v_0) = \frac{1}{2\sigma^2}\|I_0 \circ \phi_1^{-1} - I_1\|^2 + (Lv_0, v_0), \tag{2.16}$$

where (m_0, v_0) denotes the pairing of the momentum vector $m_0 \in V^*$ with the tangent vector v_0.

At an optimal solution to (2.16), the initial momenta, $m_0 = Lv_0$, is orthogonal to the level sets of the atlas (as shown in [49]). Therefore, each initial momentum m_0 is typically represented as a scalar field P multiplied by the gradient of the atlas, i.e., $m_0(x) = \nabla I(x)P(x)$. This method has a major disadvantage while solving the optimization problem: the coupled estimation of atlas and momenta leads to a poor convergence performance.

Recently, Singh et al. [57] proposed to decouple the estimation of the atlas from momenta by optimizing in the full space of vector momenta, rather than restricting it to scalar multiples of the image gradient. They demonstrated that this approach obtains better convergence rates and numerical stability. The vector momenta formulation also results in closed-form updates for optimal atlas building.

2.3.4 Diffeomorphic Atlas Building

When the energy above is minimized over all initial velocities, it yields a squared distance metric between the two input images, i.e.,

$$d(I_0, I_1)^2 = \min_{v_0 \in V} E(v_0, I_0, I_1).$$

Using this distance metric between images, the atlas estimation problem can be formulated as a least-squares estimation problem, or in other words, a Fréchet mean. Given input images $J^1, \ldots, J^N \in L^2(\Omega, \mathbb{R})$, the diffeomorphic atlas building problem is to find a template image I and initial velocities v_0^n that minimize the sum-of-squared distances function, i.e.,

$$\arg\min_I \frac{1}{N}\sum_{n=1}^{N} d(I, J^n)^2. \tag{2.17}$$

Because the distance function between images is itself a minimization problem, the atlas estimation is typically done by alternating between the minimization in (2.16) to find the optimal v_0^n and the minimization in (2.17) to update the atlas I.

Using the LDDMM distance metric between images, the atlas estimation problem can be formulated as a least-squares estimation problem, or in other words, a Fréchet mean.

As shown in Figure 2.7, given input images, a minimization of the sum-of-squared distance function is solved to estimate the atlas, $I \in L^2(\Omega, \mathbb{R})$ and the diffeomorphic transformations between the atlas and each input image as

$$E(v_0^n, I) = \sum_{n=1}^{N} \frac{1}{2\sigma^2} \left\| I \circ (\phi^n)^{-1} - J^n \right\|_{L^2}^2 + (Lv_0^n, v_0^n), \tag{2.18}$$

where the tangent vectors $\{v_0^n \in L^2([0,1], V)\}_{k=1\ldots N}$ are velocity fields in a reproducing kernel Hilbert space, V, equipped with the metric, L. The deformation ϕ^n is defined in (2.12) as the integral flow of v_0^n with $\phi_0^n = Id$. Because the distance function between images is itself a minimization problem, the atlas estimation is typically done by alternating between the minimization to find the optimal v_t^n and the update of the atlas, I. Note that for notation simplicity, we denote v_0^n as v^n in the following chapters.

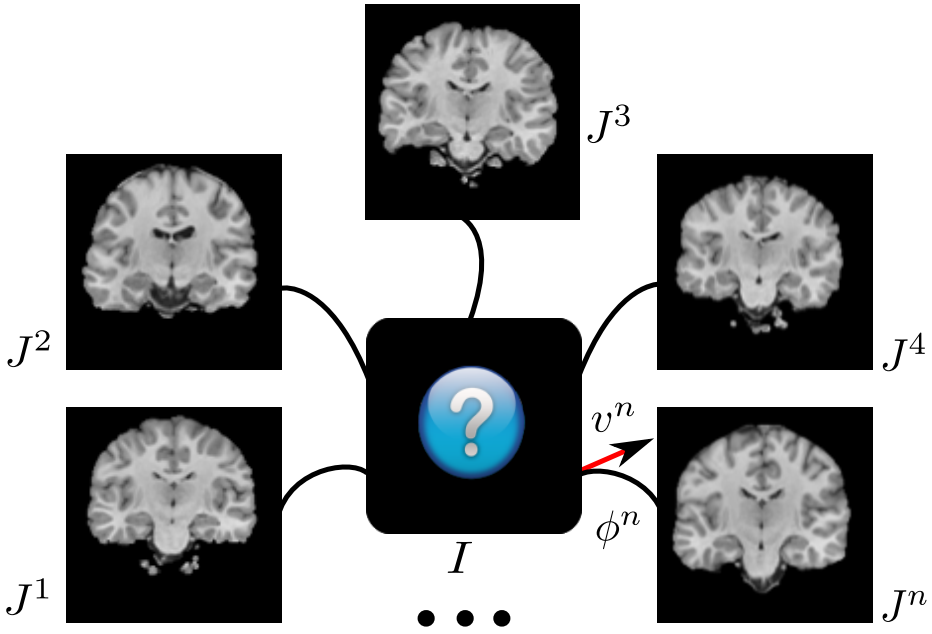


Figure 2.7: Atlas building.

CHAPTER 3

BAYESIAN ESTIMATION OF REGULARIZATION AND ATLAS BUILDING

This chapter first presents a generative Bayesian model for diffeomorphic image registration and atlas building. An atlas estimation procedure that simultaneously estimates the parameters regularizing the smoothness of the diffeomorphic transformations is developed. To achieve this, we introduce a Monte Carlo Expectation Maximization algorithm, where the expectation step is approximated via Hamiltonian Monte Carlo sampling on the manifold of diffeomorphisms. An added benefit of this stochastic approach is that it can successfully solve difficult registration problems involving large deformations, where direct geodesic optimization fails. Using synthetic data generated from the forward model with known parameters, we demonstrate the ability of our model to successfully recover the atlas and regularization parameters. We also demonstrate the effectiveness of the proposed method in the atlas estimation problem for 3D brain images.

A single atlas is not sufficiently expressive to capture distributions of images with multiple modes. This chapter later extends our single atlas estimation setting to a mixture model for building diffeomorphic multiatlases that can represent subpopulations without knowing the category of each observed data point. A key benefit of the mixture modeling inference is that it results in an automatic clustering of the dataset. Using both 2D synthetic data and 3D brain images, we show the ability of our model to successfully recover the multiatlas and automatically cluster the dataset. These models are also presented in [58] and [59] separately.

3.1 Related Work

Several works have proposed probabilistic motivations of the "groupwise" image registration problem, both in the small deformation [60], [61] and diffeomorphic [62], [63], [64] setting. In these approaches, a set of input images is registered to a template, which is

simultaneously estimated in an alternating optimization strategy. Allassonniére et al. [2] were the first to point out that atlas estimation via this alternating optimization scheme is not completely faithful to the probabilistic interpretation. They go on to propose a fully generative probability model for an image atlas and population. Later, Allassonniére et al. [65] developed a stochastic approximative expectation maximization (SAEM) algorithm to estimate the atlas and registration parameters. This estimation was done by appropriately marginalizing over the posterior distribution for the image deformations using a Monte Carlo sampling procedure.

Another related area of research involves Bayesian models of the segmentation problem. Van Leemput [66] developed a Bayesian model of the image segmentation problem that includes an atlas image and a generative deformation and image intensity model. He introduced a sampling procedure for image deformations also based on HMC, although his registration is based on a small deformation model and ours is in the diffeomorphic setting. Iglesias et al. [67] later extended this work to include uncertainty in the registration parameters by introducing hyperpriors on the parameters and integrating over their posterior. Risholm et al. [68], [69] also formulated a Bayesian model for elastic image registration and provided an MCMC method for sampling deformations, with the goal of quantifying uncertainty in the image registrations. Simpson et al. [70] furthermore inferred the level of regularization in nonrigid registration by a hierarchical Bayesian model.

Our work is the first in the diffeomorphic setting to bring MCMC sampling and correct parameter estimation via marginalization of the image transformations. Ma et al. [71] introduced a Bayesian formulation of the diffeomorphic image atlas problem, but also estimated the atlas using a mode approximation to alternate between atlas and registration optimizations. They do not estimate the registration parameters. There has been some work on stochastic flows of diffeomorphisms [72], which are Brownian motions, i.e., small perturbations integrated along a time-dependent flow. This differs from the prior distribution in our work, which is on the tangent space of initial velocity fields, rather than on the entire time-dependent flow. Our formulation leads to random geodesics in the space of diffeomorphisms, and makes possible an efficient sampling procedure for MCMC sampling.

3.2 A Bayesian Model for Diffeomorphic Atlas Building

For a continuous domain Ω, direct interpretation of the image match term as a negative log posterior is problematic, as it would be akin to isotropic Gaussian noise in the infinite-dimensional Hilbert space $L^2(\Omega, \mathbb{R})$. This is not a well-defined probability distribution as

it has infinite measure. More appropriately, we can instead consider our input images, J^n, and our atlas image, I, to be measured on a discretized grid Ω. That is, images are elements of the finite-dimensional Euclidean space $L^2(\Omega, \mathbb{R})$. We will also consider velocity fields v^n and the resulting diffeomorphisms ϕ^n to be defined on the discrete grid, Ω. Now our noise model is i.i.d. Gaussian noise at each image voxel, with likelihood given by

$$p(J^n \mid v^n, I) = \frac{1}{(2\pi)^{M/2}\sigma^M} \exp\left(-\frac{\|I \circ (\phi^n)^{-1} - J^n\|^2}{2\sigma^2}\right), \tag{3.1}$$

where M is the number of voxels, σ^2 is the noise variance, and the norm inside the exponent is the Euclidean norm of $L^2(\Omega, \mathbb{R})$.

The negative log prior on the v^n is a discretized version of the squared Hilbert space norm above. Now consider L to be a discrete, self-adjoint, positive-definite differential operator on the domain Ω. The prior on each v^n is given by a multivariate Gaussian,

$$p(v^n) = \frac{1}{(2\pi)^{\frac{M}{2}}|L^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{(Lv^n, v^n)}{2}\right), \tag{3.2}$$

where d is the dimension of v^n, and $|L|$ is the determinant of L. In this work, we use a metric of the form $L = -\alpha\Delta + \beta$, where Δ is the discrete Laplacian, and α and β are positive numbers. In the sequel, we consider $\theta = (\alpha, \sigma, I)$ to be parameters that we wish to estimate. We fix β to a small number to ensure that the L operator is nonsingular. Putting together the likelihood (3.1) and prior (3.2), we arrive at the log joint posterior for the diffeomorphisms, via initial velocities, v^n, as

$$\begin{aligned} \log \prod_{n=1}^{N} p\left(v^n \mid J^n; \theta\right) \propto\ & \frac{N}{2}\log|L| - \frac{1}{2}\sum_{n=1}^{N}(Lv^n, v^n) \\ & - \frac{MN}{2}\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\|I \circ (\phi^n)^{-1} - J^n\|^2. \end{aligned} \tag{3.3}$$

3.3 Estimation of Model Parameters

We now present an algorithm for estimating the parameters, θ, of the probabilistic image atlas model specified in the previous section. These parameters include the image atlas, I, the smoothness level, or metric parameter, α, and the standard deviation of the image noise, σ. We treat the v^n, i.e., the initial velocities of the image diffeomorphisms, as latent random variables with log posterior given by (3.3). This requires integration over the latent variables, which is intractable in closed form. We thus develop a Hamiltonian Monte Carlo procedure for sampling v^n from the posterior and use this in a Monte Carlo Expectation Maximization algorithm to estimate θ. This procedure consists of two main steps:

1. E-step We draw a sample of size S from the posterior distribution (3.3) using HMC with the current estimate of the parameters, $\theta^{(i)}$. Let v^{nj}, $j = 1, \ldots, S$, denote the jth point in this sample for the kth velocity field. The sample mean is taken to approximate the Q function,

$$\begin{aligned} Q(\theta \,|\, \theta^{(i)}) &= E_{v^n|J^n;\theta^{(i)}} \left[\sum_{n=1}^{N} \log p\left(v^n \,|\, J^n; \theta\right) \right] \\ &\approx \frac{1}{S} \sum_{j=1}^{S} \sum_{n=1}^{N} \log p\left(v^{nj} \,|\, J^n; \theta\right). \end{aligned} \tag{3.4}$$

2. M-step Update the parameters by maximizing $Q(\theta \,|\, \theta^{(i)})$. The maximization is closed form in I and σ, and a one-dimensional gradient ascent in α.

In the HMC sampling procedure, we need to compute gradients, with respect to initial momenta $m^n = Lv^n$, of the diffeomorphic image matching problem in (2.16), for matching the atlas I to an input image J^n. Following the optimal control theory approach in [1], [57], we add Lagrange multipliers to constrain the diffeomorphism $\phi^n(t)$ being a geodesic path. After introducing time-dependent adjoint variables, $\hat{m}, \hat{I}$, and $\hat{v}$, we write the augmented energy as,

$$\tilde{E}(m^n) = E(Km^n, I, J^n) + \int_0^1 \langle \hat{m}, \dot{m}^n + \mathrm{ad}^*_{v^n} m^n \rangle + \langle \hat{I}, \dot{I} + \nabla I \cdot v^n \rangle + \langle \hat{v}^n, m^n - Lv^n \rangle dt,$$

where E is the diffeomorphic image matching energy from (2.16), and the other terms correspond to Lagrange multipliers enforcing a) the geodesic constraint, which comes from the EPDiff equation (2.13), b) the image transport equation, $\dot{I} = -\nabla I \cdot v^n$, and c) the constraint that $m^n = Lv^n$, respectively.

The optimality conditions for m^n, I, v^n are given by the following time-dependent system of ODEs, termed the *adjoint equations*:

$$-\dot{\hat{m}} + \mathrm{ad}_{v^n} \hat{m} + \hat{v} = 0, \qquad -\dot{\hat{I}} - \nabla \cdot (\hat{I} v^n) = 0, \qquad -\mathrm{ad}^*_{\hat{m}} m^n + \hat{I} \nabla I - L\hat{v} = 0,$$

subject to initial conditions

$$\hat{m}(1) = 0, \qquad \hat{I}(1) = \frac{1}{\sigma^2}(I(1) - J^n).$$

Finally, after integrating these adjoint equations backwards in time to $t = 0$, the gradient of $\tilde{E}$ with respect to the initial momenta is

$$\nabla_{m^n} \tilde{E} = Km^n - \hat{m}(0). \tag{3.5}$$

3.3.1 Hamiltonian Monte Carlo (HMC) Sampling

Hamiltonian Monte Carlo [73] is a powerful MCMC sampling methodology that is applicable to a wide array of continuous probability distributions. It utilizes Hamiltonian dynamics as a Markov transition probability and efficiently explores the space of a target distribution. The integration through state space results in more efficient, global moves, while it also uses gradient information of the log probability density to sample from higher probability regions. In this section, we derive an HMC sampling method to draw a random sample from the posterior distribution of our latent variables, v^n, the initial velocities defining the diffeomorphic image transformations from the atlas to the data.

To sample from a pdf $f(x)$ using HMC, one first sets up a Hamiltonian $H(x, \mu) = U(x) + V(\mu)$, consisting of a "potential energy", $U(x) = -\log f(x)$, and a "kinetic energy", $V(\mu) = -\log g(\mu)$. Here $g(\mu)$ is some proposal distribution (typically isotropic Gaussian) on an auxiliary momentum variable, μ. An initial random momentum μ is drawn from the density $g(\mu)$. Starting from the current point x and initial random momentum μ, the Hamiltonian system is integrated forward in time to produce a candidate point, $\tilde{x}$, along with the corresponding forward-integrated momentum, $\tilde{\mu}$. The candidate point $\tilde{x}$ is accepted as a new point in the sample with probability

$$P(\text{accept}) = \min(1, \exp(-U(\tilde{x}) - V(\tilde{\mu}) + U(x) + V(\mu)).$$

This acceptance-rejection method is guaranteed to converge to the desired density $f(x)$ under fairly general regularity assumptions on f and g.

In our model, to sample v^n from the posterior in (3.3), we equivalently sample m^n from the dual momenta, using $v^n = Km^n$, so we define our potential energy as

$$U(m^n) = -\log p(m^n | J^n; \theta).$$

We use the prior distribution on the dual momenta as our proposal density, in other words, we use $p(K\mu)$ defined as in (3.2), taking care to include the appropriate change-of-variables. This gives the kinetic energy, $V(\mu) = (\mu, K\mu)$. This gives us the following Hamiltonian system to integrate in the HMC:

$$\begin{aligned} \frac{dm^n}{dt} &= \frac{\partial H}{\partial \mu} = K\mu, \\ \frac{d\mu}{dt} &= -\frac{\partial H}{\partial m^n} = -\nabla_{m^n} \tilde{E}, \end{aligned}$$

where the last term comes from the gradient defined in (3.5). As is standard practice in HMC, we use a "leap-frog" integration scheme, which better conserves the Hamiltonian and results in high acceptance rates.

3.3.2 The Maximization Step

We now derive the M-step for updating the parameters $\theta = (\alpha, \sigma, I)$ by maximizing the HMC approximation of the Q function, which is given in (3.11). This turns out to be a closed-form update for the noise variance σ^2 and the atlas I, and a simple one-dimensional gradient ascent for α.

From (3.3) and (3.11), it is easy to derive the closed-form update for σ as

$$\sigma^2 = \frac{1}{MNS} \sum_{j=1}^{S} \sum_{n=1}^{N} \|I_0 \circ (\phi^{nj})^{-1} - J^n\|^2. \tag{3.6}$$

For updating the atlas image I, we set the derivative of the Q function approximation with respect to I to zero. The solution for I gives a closedform update,

$$I = \frac{\sum_{j=1}^{S} \sum_{n=1}^{N} J^n \circ \phi^{nj} |D\phi^{nj}|}{\sum_{j=1}^{S} \sum_{n=1}^{N} |D\phi^{nj}|}.$$

The gradient ascent over α requires that we take the derivative of the metric $L = -\alpha\Delta + \beta I$, with respect to α. We do this in the Fourier domain, where the discrete Laplacian is a diagonal operator. For a 3D grid, the coefficients A_{xyz} of the discrete Laplacian at coordinate (x, y, z) in the Fourier domain are

$$A_{xyz} = -2 \left(\cos \frac{2\pi x}{W} + \cos \frac{2\pi y}{H} + \cos \frac{2\pi z}{D} \right) + 6,$$

where W, H, D are the dimension of each direction. Hence, the determinant of the operator L is

$$|L| = \prod_{x,y,z} A_{xyz}\alpha + \beta.$$

The gradient of the HMC approximated Q function, with respect to α, is

$$\nabla_\alpha Q(\theta \,|\, \theta^{(i)}) \approx \frac{1}{2} \sum_{j=1}^{S} \sum_{n=1}^{N} \left[\sum_{x,y,z} \frac{A_{xyz}}{A_{xyz}\alpha + \beta} - \left\langle -\Delta v^{nj}, v^{nj} \right\rangle \right].$$

3.4 Results

We demonstrate the effectiveness of our proposed model and MCEM estimation routine using both 2D synthetic data and real 3D MRI brain data. Because we have a generative model, we can forward simulate a random sample of images from a distribution with known parameters $\theta = (\alpha, \sigma, I)$. Then, in the next subsection, we test if we can recover those parameters using our MCEM algorithm. Figure 3.1 illustrates this process. We simulated a 2D synthetic dataset starting from an atlas image, I, of a binary circle with resolution 100×100. We then generated 20 smooth initial velocity fields from the prior distribution,

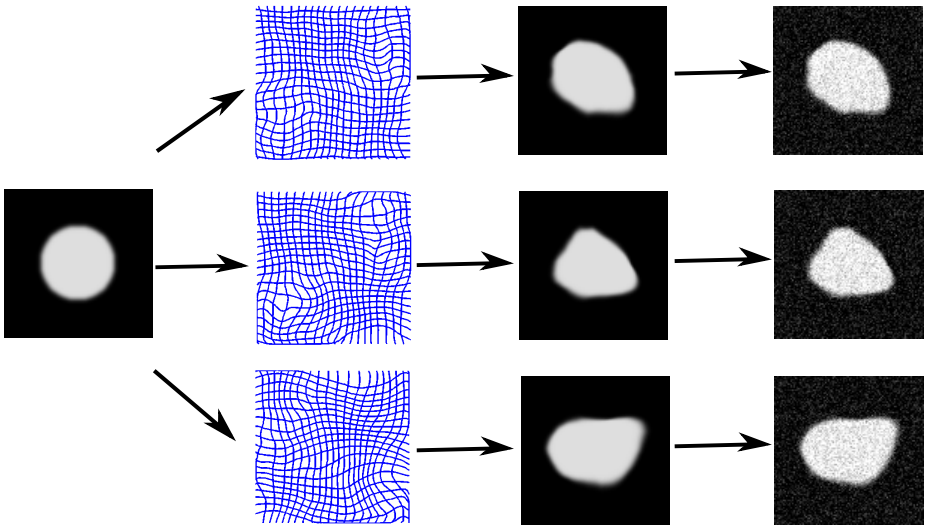


Figure 3.1: Simulating synthetic 2D data from the generative diffeomorphism model. From left to right: the ground truth template image, random diffeomorphisms from the prior model, deformed images, and final noise-corrupted images.

$p(v^n)$, defined in (3.2), setting $\alpha = 0.025$ and $\beta = 0.001$. Deformed circle images were constructed by shooting the initial velocities by the EPDiff equations and transforming the atlas by the resulting diffeomorphisms, ϕ^n. Finally, we added i.i.d. Gaussian noise according to our likelihood model (3.1). We used a standard deviation of $\sigma = 0.05$, which corresponds to an SNR of 20 (which is more noise than typical structural MRI).

3.4.1 Parameter Estimation on Synthetic Data

In our estimation procedure, we initialized α with 0.002 for noise-free and 0.01 for noise-corrupted images. The step size of 0.005 with 20 steps for leap-frog integration is used in HMC with 10 units of time discretization in integration of EPDiff equations.

Figure 3.2 compares the true atlas and estimated atlases in the clean and noisy case. Figure 3.3 shows the convergence graph for α and σ estimation by using 100 samples with another 50 burn in. It shows that our method recovers the model parameters fairly well. However, the iterative mode approximation algorithm does not recover the α parameter as nicely as our method. In the noisy case, the mode approximation algorithm estimates α as 0.0152, which is far from the ground truth value of 0.025. This is compared with our estimation of 0.026. In addition, in the noise-free example, the mode approximation algorithm blows up due to the σ dropping close to 0, thus making the image match term numerically too high and the geodesic shooting unstable.

Figure 3.2: Atlas estimation results. Left: ground-truth template. Center: estimated template from noise-free dataset. Right: estimated template from noise-corrupted dataset.

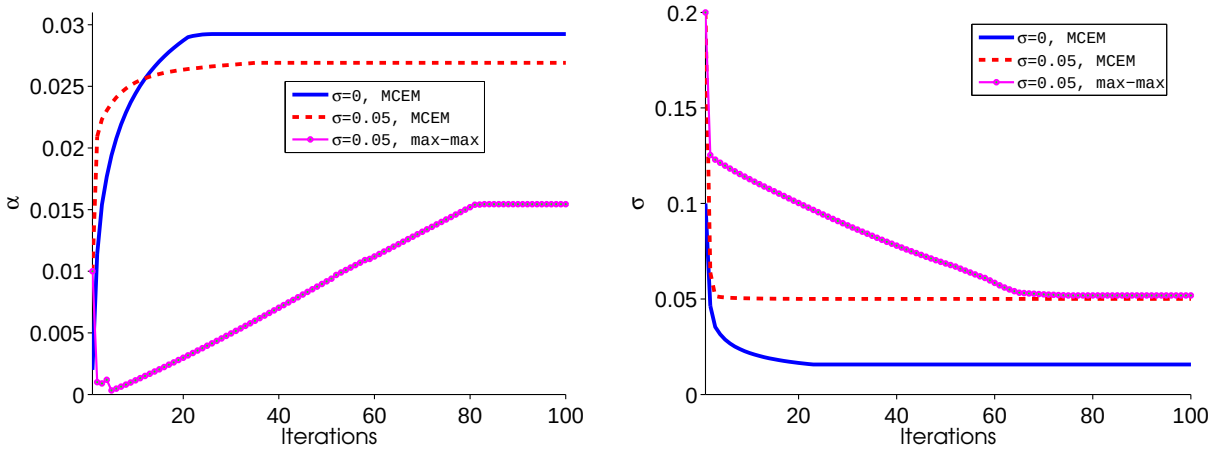


Figure 3.3: Estimation of α, σ. Left: α estimation. Right: σ estimation. In our MCEM method, final estimated α and σ for noise-free data are 0.028, 0.01, and for noise data are 0.026, 0.0501. Compared with max-max method, for the noise data, estimated α and σ are 0.0152, 0.052.

3.4.2 Atlas Building on 3D Brain Images

To demonstrate the effectiveness of our method on the real data, we apply our MCEM atlas estimation algorithm to a set of brain MRI from 10 healthy subjects. The MRI have resolution $108 \times 128 \times 108$ and are skull-stripped, intensity normalized, and co-registered with rigid transforms. We set the initial $\alpha = 0.01$, $\beta = 0.001$ with 15 time-steps.

The left side of Figure 3.4 shows coronal and axial slices from the 3D MRI used as input. The right side shows the initialization (grayscale average of the input images), followed by the final atlas estimated by our method. The final atlas estimate correctly aligns the anatomy of the input images, producing a sharper average image. The algorithm also jointly estimated the smoothness parameter to be $\alpha = 0.028$ and the image noise standard deviation to be $\sigma = 0.031$.

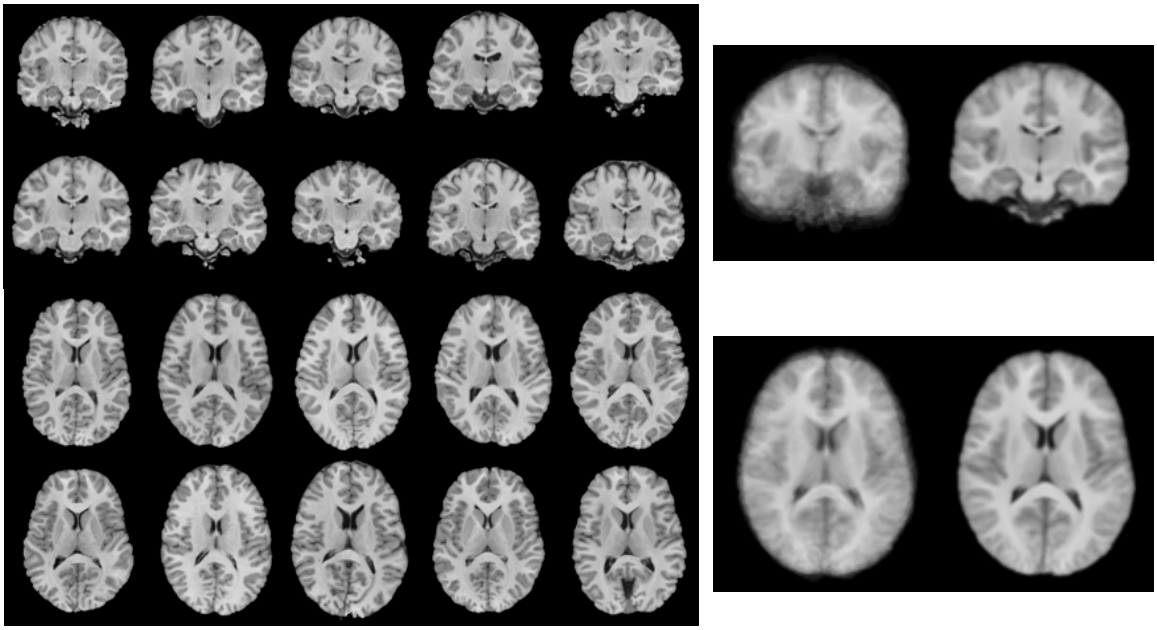


Figure 3.4: Left: coronal and axial slices from the input 3D MRIs. Middle: initial grayscale average of the input images. Right: final atlas estimated by our MCEM estimation procedure.

We compare the estimated atlas and one of the deformed subject at different levels of α at 2.8, $2.8e-1$, $2.8e-2$, $2.8e-3$, $2.8e-4$, where $\alpha = 2.8e-2$ is estimated by our model (see Figure 3.5). The figure shows how α affects the estimation of atlas and deformations. If α gets too high, which means too much smoothness and not enough deformation, the atlas is blurred. Otherwise, nonsmooth deformation introduces artifacts in both atlas and deformed images.

3.4.3 Image Matching Accuracy

Finally, we demonstrate that another benefit of our HMC sampling methodology is improved performance in the standard image registration problem under large deformation shooting. Rather than use a direct gradient descent to solve the image registration problem, we instead can find the posterior mean of the model (3.3), where for image matching we fix the "atlas", I, as the source image and have just one target image, I_1. The stochastic behavior in the sampling helps to get out of local minimum, where the direct gradient descent can get stuck. We compared our proposed method with direct gradient descent image registration by geodesic shooting from [1]. We used the authors' uTIlzReg package for geodesic shooting, which is available freely online. For the comparison, we registered the image pair shown in the first two panels of Figure 3.6, which requires a large deformation.

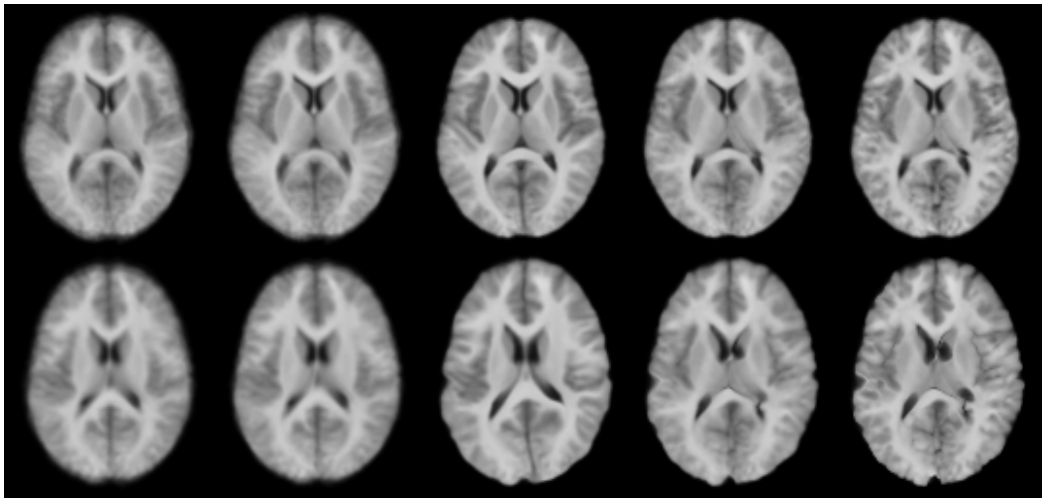


Figure 3.5: Top to bottom: estimated atlas and one of the deformed subject. Left to right: comparison of different value of α at 2.8, $2.8e-1$, $2.8e-2$, $2.8e-3$, $2.8e-4$, where $\alpha = 2.8e-2$ is our estimation.

The source and target images are 50×50. We used $\alpha = 0.02, \beta = 0.001$ for smoothing kernel, and $h = 40$ time-steps between $t = 0$ and $t = 1$. Note that we only want to compare the image matching here, so we fix the α and σ parameters.

Figure 3.6 demonstrates the results of the direct geodesic shooting registration with our HMC posterior mean. It shows that the geodesic shooting method gets stuck in a local minimum and cannot make it to the target image even with a large number of time-steps ($h = 60$) in the time discretization (we tried several time discretizations up to 60, and none worked). Though our method did not match perfectly in the tip of the "C", it still recovers the full shape while retaining a diffeomorphic transformation.

Figure 3.6: The first two images from left to right are the source and target image respectively. The third is the matched image obtained by geodesic shooting method using [1]. The Last image is the matched image from our MCEM method.

3.5 Mixture Model of Diffeomorphic Multiatlas Building

A single atlas does not provide enough information for robust statistical analysis if significant differences exist between subpopulations. Blezek et al. [74] were the first to investigate the multiatlas building problem and infer each atlas from the mode of a population through the mean shift algorithm. They iteratively optimized between a small deformation image registration framework and atlas construction. Later, two major classes of multiatlas building methods were developed in medical imaging. One of the classes is a two-step strategy, in which the algorithm does clustering, such as K-means or affinity propagation, after registration. Another class of multiatlas building is motivated by probabilistic modeling of multiatlases. Allassonnière et al. [2] discussed a mixture model of template estimation with small deformations. Sabuncu et al. [75] introduced a joint framework of image registration and clustering using a mixture of Gaussians, although their work was not in a diffeomorphic setting. Tang et al. [76] proposed a random diffeomorphic orbit model to treat the multiple atlases as Gaussian random fields, and then estimated them from the model using maximum a posteriori estimation. These multiatlas methods are of high importance for related research areas, for example, image segmentation [77], [79], [80], where a priori knowledge about the shapes and structures from the presegmented multiple atlases is used to guide the segmentation [78], [81]. Aljabar et al. [77] discussed the issue of multiatlas selection and showed that multiatlas based segmentation results in higher accuracy than a single atlas.

In this section, we present a mixture model for building diffeomorphic multiatlases that can represent subpopulations without knowing the category of each observed data point. This work can for the first time cluster population-based images into different subgroups automatically while co-registering them in a diffeomorphic setting with marginalized deformations. We again treat diffeomorphic image transformations as latent variables and integrate them out from the posterior distribution.

3.5.1 Our Mixture Model

We assume that the input images $\{J^n\}_{n=1,\ldots,N}$ are generated from multiple atlases I^k, where $k = 1, \ldots, K$ represents the number of clusters and π^k denotes the prior probability of the kth cluster. Each individual image J^n is associated with a k-dimensional binary random variable $\boldsymbol{z}^n$, in which a particular kth element z^{nk} is equal to 1 and all other elements are equal to 0. As in a general mixture model, the prior distribution of $\boldsymbol{z}^n$ is specified by the mixing coefficients π^k as $p(z^{nk} = 1) = \pi^k$, where $\pi^k \in [0, 1]$ with $\sum_{k=1}^{K} \pi^k = 1$. We then can write the distribution of $p(\boldsymbol{z}^n)$ as

$$p(\boldsymbol{z}^n) = \prod_{k=1}^{K} (\pi^k)^{z^{nk}}.$$

Let $\boldsymbol{v}^n$ denote a set of initial velocities from each cluster k for the nth image, which is $\{v^{nk}\}$. Similarly, the atlases $\{I^k\}$ and noise variances $\{\sigma^k\}$ will be represented as $\boldsymbol{I}$ and $\boldsymbol{\sigma}$. Consider that our input images and atlases are measured on a discrete grid, we formulate our noise model as i.i.d. Gaussian at each image voxel, with the data likelihood $p(J^n \mid \boldsymbol{z}^n, \boldsymbol{v}^n, \boldsymbol{I}, \boldsymbol{\sigma})$ given by

$$\begin{aligned} p(J^n \mid \boldsymbol{z}^n, \boldsymbol{v}^n, \boldsymbol{I}, \boldsymbol{\sigma}) &= \prod_{k=1}^{K} N(J^n \mid v^{nk}, I^k, \sigma^k)^{z^{nk}} \\ &= \prod_{k=1}^{K} \left[\frac{1}{(2\pi)^{M/2}(\sigma^k)^M} \exp\left(-\frac{\|I^k \circ (\phi^{nk})^{-1} - J^n\|^2}{2(\sigma^k)^2} \right) \right]^{z^{nk}}, \end{aligned} \tag{3.7}$$

where M is the number of voxels.

We then define a multivariate Gaussian distribution on the initial velocity $\boldsymbol{v}^n$ that guarantees smoothness of the geodesic shooting path. The formulation is given by

$$\begin{aligned} p(\boldsymbol{v}^n) &= \prod_{k=1}^{K} N(v^{nk} \mid 0,\, L^{-1})^{z^{nk}} \\ &\propto \prod_{k=1}^{K} \left[\exp\left(-\frac{1}{2}(Lv^{nk}, v^{nk}) \right) \right]^{z^{nk}}, \end{aligned} \tag{3.8}$$

where we use a metric of the form $L = -\alpha\Delta + I$, in which Δ is the discrete Laplacian operator, α is a positive regularity parameter, and I denotes an identity matrix. In this chapter, we set α with the same value across all clusters.

Putting together equations (3.7) and (3.8), we arrive at the log joint posterior distribution with a set of parameters $\theta = \{I^k, \sigma^k, \pi^k\}$ as

$$\begin{aligned} \log \prod_{n=1}^{N} p(\boldsymbol{z}^n, \boldsymbol{v}^n \mid J^n, \theta) \;=\; & \sum_{n=1}^{N} \sum_{k=1}^{K} z^{nk} \{ \log \pi^k - \frac{1}{2(\sigma^k)^2} \left\| J^n - I^k \circ (\phi^{nk})^{-1} \right\|_{L^2}^2 \\ & - \frac{MN}{2} \log \sigma^k - \frac{1}{2}(Lv^{nk}, v^{nk}) \} + const. \end{aligned} \tag{3.9}$$

Figure 3.7 shows the graphical representation of our model.

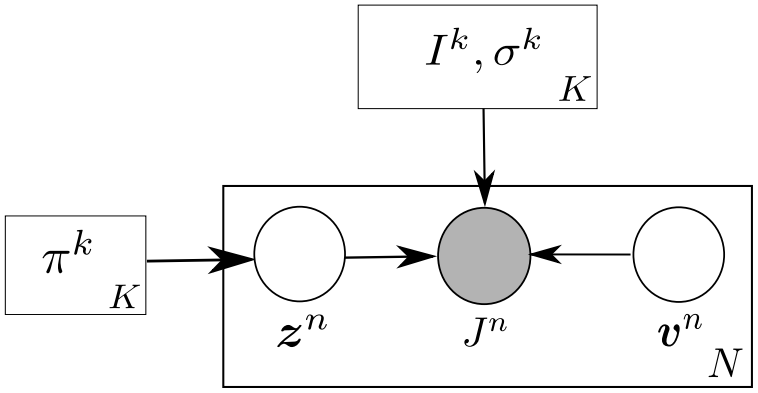


Figure 3.7: Graphical representation of our model for a set of i.i.d. images $\{J^n\}$, with corresponding latent variables $\{\boldsymbol{z}^n, \boldsymbol{v}^n\}_{n=1,\ldots,N}$ and parameters $\{I^k, \sigma^k, \pi^k\}_{k=1,\ldots,K}$.

3.6 Inference

Similar to single atlas building, we use MCEM to infer and estimate the parameters $\theta = \{I^k, \sigma^k, \pi^k\}$. In order to treat the $\boldsymbol{z}^n, \boldsymbol{v}^n$ as latent random variables, we need to integrate them out over the log posterior given by (3.9). Marginalizing $\boldsymbol{z}^n$ is straightforward as the Gaussian mixture model. However, marginalizing $\boldsymbol{v}^n$ is intractable in the closed form. We generate samples $\boldsymbol{v}^n$ from the posterior distribution (3.9), and use these samples in a Monte Carlo Expectation Maximization algorithm to estimate θ. The inference consists of two main steps:

3.6.1 E-step

To compute the expectation function Q, we integrate out the hidden variables $\boldsymbol{z}^n$ and $\boldsymbol{v}^n$ with the current estimate of the parameters $\theta^{(i)}$ as

$$Q(\theta \mid \theta^{(i)}) = E_{\boldsymbol{z}^n, \boldsymbol{v}^n \mid \theta^{(i)}} \left[\sum_{n=1}^{N} \log p(\boldsymbol{z}^n, \boldsymbol{v}^n \mid J^n, \theta) \right]. \tag{3.10}$$

A standard way to approximate (3.10) is sampling $\boldsymbol{z}^n, \boldsymbol{v}^n$ through Gibbs sampling on the joint posterior distribution (3.9). We draw S samples, ${\boldsymbol{v}^{nj}}_{\{j=1,\ldots,S\}}$, from the log conditional distribution $\log p(\boldsymbol{v}^n \mid \boldsymbol{z}^n, J^n, \theta^{(i)})$ by HMC. Note that $\boldsymbol{v}^{nj}$ denotes a set of jth samples for the nth initial velocity across k clusters. We then use the sample mean to approximate the expectation function Q. To simplify the computation, we develop a closed-form solution for marginalizing $\boldsymbol{z}^n$ from the log conditional distribution $\log p(\boldsymbol{z}^n \mid \boldsymbol{v}^n, J^n, \theta^{(i)})$ directly.

Much like [82], the closed-form solution for computing the expectation of z^{nk} is

$$\gamma(z^{nk}) = \frac{\pi^k \prod_{j=1}^{S} N(J^n, v^{njk} \mid \theta^{(i)})}{\sum_{k=1}^{K} \pi^k \prod_{j=1}^{S} N(J^n, v^{njk} \mid \theta^{(i)})},$$

which is the responsibility that cluster k takes for representing the observed image data J^n. Here v^{njk} is the jth sample for the nth velocity field that belongs to the cluster k.

The final expectation function (3.10) is ultimately approximated as

$$Q(\theta \,|\, \theta^{(i)}) \approx \frac{1}{S} \sum_{j=1}^{S} \sum_{n=1}^{N} \log p(\gamma(\boldsymbol{z}^n), \boldsymbol{v}^{nj} \,|\, J^n, \theta^{(i)}). \tag{3.11}$$

3.6.2 M-step

We then maximize the approximated function $Q(\theta \,|\, \theta^{(i)})$ (3.11) to update the parameters $\theta = \{I^k, \sigma^k, \pi^k\}$, which turns out to be a closed-form update for all parameters. We set the derivative of the expectation Q w.r.t. each parameter of θ to zero, and the closed-form update is

$$\begin{aligned}
\tilde{\pi}^k &= \frac{N^k}{N}, \text{ where } N^k = \sum_{n=1}^{N} \gamma(z^{nk}), \\
\tilde{I}^k &= \frac{\sum_{n=1}^{N} \sum_{j=1}^{S} \gamma(z^{nk}) \cdot J^k \circ \phi^{njk} |D\phi^{njk}|}{\sum_{n=1}^{N} \sum_{j=1}^{S} \gamma(z^{nk}) \cdot |D\phi^{njk}|}, \\
(\tilde{\sigma}^k)^2 &= \frac{1}{M \cdot S \cdot N^k} \sum_{i=1}^{N} \sum_{j=1}^{S} \gamma(z^{nk}) \cdot \| I^k \circ (\phi^{njk})^{-1} - J^n \|^2.
\end{aligned}$$

3.7 Results

We demonstrate the effectiveness of our model using both 2D synthetic data and real 3D MRI brain data.

3.7.1 Synthetic Data

Because we have a generative model, we can forward simulate random images from the known parameters $\theta = \{I^k, \sigma^k, \pi^k\}$, where we choose $k \in \{1, 2, 3\}$. We use three atlases, which are 2D binary images of a square, triangle, and ellipse with a resolution of 100×100. We then generate 30 initial velocity fields (10 per cluster) from the prior $p(v^{nk})$ given in (3.8), setting $\alpha = 3.0$. We shoot the initial velocities by the EPDiff equations (2.13) to generate diffeomorphic deformations, and then use them to transform the atlases. Finally, we add random Gaussian noise with $\sigma = (0.01, 0.025, 0.03)$ to each transformed cluster atlas.

In our testing procedure, we initialize $\sigma = 0.3$ for all K clusters (K is the true number of clusters in this synthetic example). For the HMC sampling procedure, we use a step size of 0.05 for leap-frog integration with 40 samples after a burn-in of 50 samples. Each atlas from the kth cluster is initialized to the linear average of the image intensities over

the samples we generated for each cluster, and the $\{\pi^k\}$ are set as the averaged weight, 0.3. Figure 3.8 compares the ground truth atlases and our estimated atlases, showing that our model is able to accurately recover the true atlases, as well as automatically cluster data into subgroups. As for other parameters, we get the estimated $\sigma = (0.011, 0.026, 0.031)$ and $\pi = 0.33333$ for each cluster. We compared our multiatlas approach with a single atlas estimated over all data points using the method of Zhang et al. [58]. The single atlas was completely incapable of representing the synthetic dataset.

3.7.2 OASIS Brain Data

To show the effectiveness of our model on the real 3D brain data, we applied our algorithm to an OASIS brain MRI dataset with 26 healthy subjects from ages 60 to 90. All the MRI images have resolution $128 \times 128 \times 128$ with the image spacing $1.0 \times 1.0 \times 1.0mm^3$, and are skull-stripped, intensity normalized, and co-registered with affine transforms. We set $\alpha = 0.3$, which was estimated by the single atlas building framework with 10 time-steps in geodesic shooting. We ran K-means algorithm with two clusters using image intensity as features, and then used the output as our initialization for I^k, the initial atlas at each cluster. Note that here we use cross-validation to determine the number of clusters. Other alternative ways could also be used, such as the Elbow method, which evaluates the percentage of variance with respect to the number of clusters and information criterion approaches (for instance, Akaike information criterion and Bayesian information criterion).

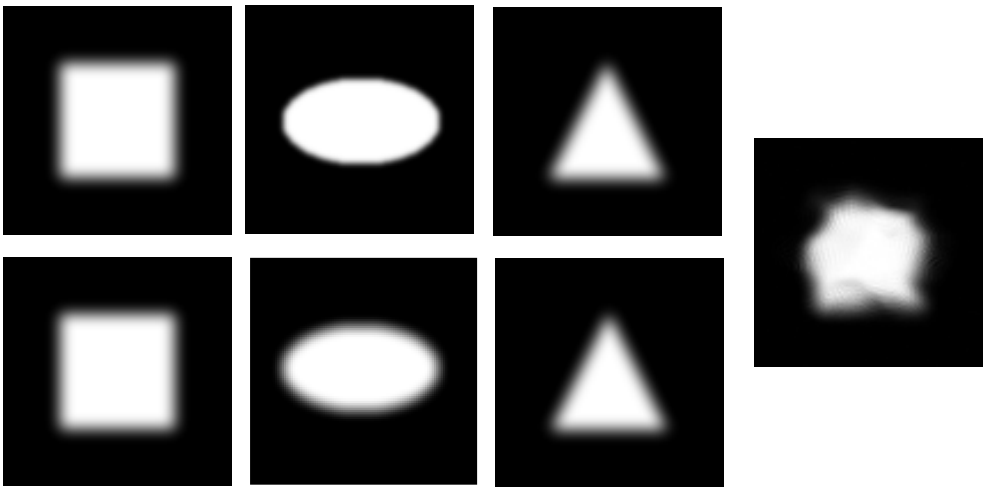


Figure 3.8: Estimation of atlases. Top: ground truth atlases of three clusters: square, ellipse, and triangle; Bottom: our estimation; Right: single atlas estimated from the whole dataset.

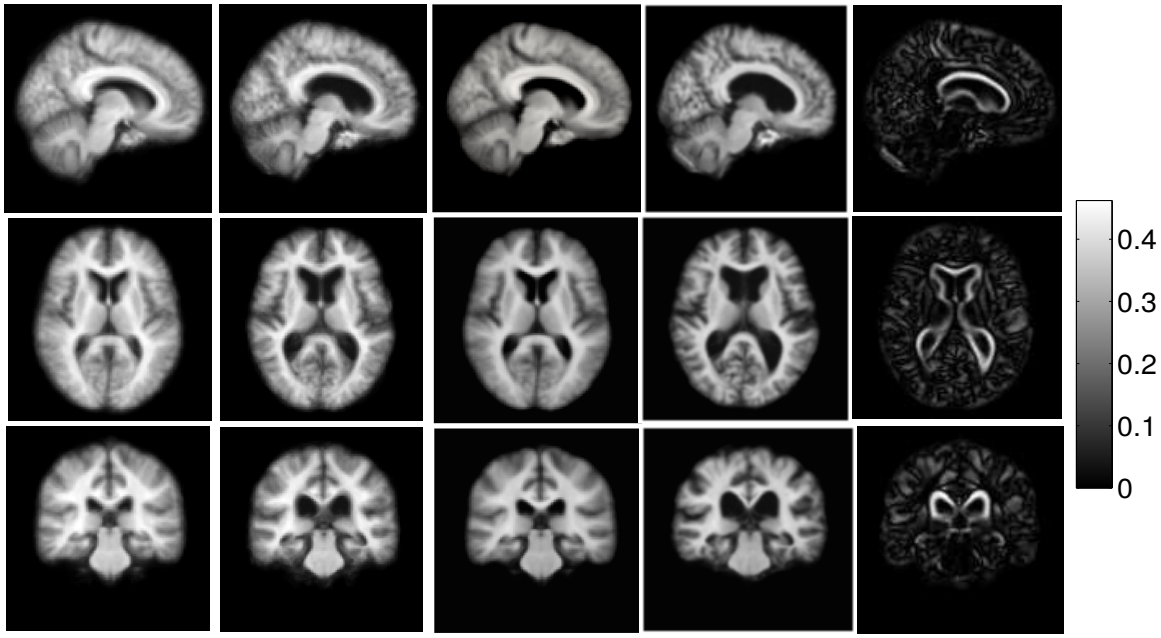


Figure 3.9: Initialization and our estimation of atlases. Top to bottom: sagittal, axial, and coronal views of the K-means initialization and our estimated atlases. Left to right: initialization for each cluster (column 1-2), our estimated atlases from two different clusters (column 3-4) and difference maps over image intensity between two atlases.

The first two columns in Figure 3.9 show sagittal, axial, and coronal views of slices from the output of K-means algorithm, which are the grayscale averages of the clustered images. The middle two columns are atlases estimated from our model. It demonstrates that the final atlases produce sharper averaged images with more details. Meanwhile, the big shape difference between the two estimated atlases shows that multiple atlases gives a better representation of the multimodel population by an atlas per mode than a single atlas that mixes up the features across different groups. For the purpose of better visualization, we also add difference maps that represent the absolute value of the intensity differences between our two estimated atlases.

3.8 Conclusion

This chapter first presented a novel generative model of the diffeomorphic atlas estimation problem. This method is the first to jointly estimate the regularity parameter, noise variance, and image atlas. It faithfully treats the diffeomorphic transformations from the atlas to the input images as unobserved random variables. We introduced a MCMC sampling scheme to integrate over these transformations. While we chose a particular parameterized form for the metric operator L, other metrics are also possible in our framework. This work

opens up the possibility of extensions for rigorous probabilistic modeling of shape variability through diffeomorphisms.

Based on the Bayesian setting of a single atlas building, this chapter then introduced a generative Gaussian mixture model of diffeomorphic multiatlas building. This is the first probabilistic model for constructing multiple atlases navigated by unsupervised clustering in a diffeomorphic setting. Our algorithm aggregates data that belongs to the same category automatically and constructs multiple representations of a large image database. This framework can be very useful for further statistical analysis in many areas, such as shape variation quantification and guidance of image segmentation.

CHAPTER 4

PROBABILISTIC PRINCIPAL GEODESIC ANALYSIS

The previous chapter introduces a Bayesian model for estimating a mean of the dataset. However, it does not encode the data variability of a population. This chapter describes the second contribution of this dissertation, which is also presented in [53], [83]. It describes a latent variable model for principal geodesic analysis (PGA) that provides a probabilistic framework for factor analysis on finite-dimensional manifolds.

This chapter first begins with a brief introduction on current related works about density estimation on manifolds and other principal modes analysis approaches except PGA.

4.1 Related Work

There has been some work on density estimation on Riemannian manifolds. For example, there is a wealth of literature on parametric density estimation for directional data [9], e.g., spheres, projective spaces, etc. Nonparametric density estimation based on kernel mixture models [84] was proposed for compact Riemannian manifolds. Methods for sampling from manifold-valued distributions have also been proposed [58], [85]. It is important to note the distinction between manifold data, where the manifold representation is known as *a priori*, versus manifold learning and nonlinear component analysis [26], [86], where the data lie in Euclidean space on some unknown, lower-dimensional manifold that must be learned.

While [27] originally proposed an approximate estimation procedure for PGA, recent contributions [87], [88] have developed algorithms for exact solutions to PGA. Related work on manifold component analysis has introduced variants of PGA, including relaxing the constraint that geodesics pass through the mean of the data [89] and, for spherical data, replacing geodesic subspaces with nested spheres of arbitrary radius [90]. All these methods are based on geometric, least-squares estimation procedures, i.e., they find subspaces that minimize the sum-of-squared geodesic distances to the data. Much like the original

formulation of PCA, current component analysis methods on manifolds lack a probabilistic interpretation.

In this chapter, we propose a latent variable model for PGA, called probabilistic PGA (PPGA). We then extend this model to a Bayesian formulation that automatically selects the intrinsic data dimensionality. The model definition applies to generic manifolds. However, due to the lack of an explicit formulation for the normalizing constant, our estimation is limited to *symmetric spaces*, which include many common manifolds such as Euclidean space, spheres, Kendall shape spaces, Grassman/Stiefel manifolds, and more.

4.2 Probabilistic Principal Geodesic Analysis

Following [34], [91], we use a generalization of the normal distribution for a Riemannian manifold as our noise model. Consider a random variable y taking values on a Riemannian manifold M, defined by the probability density function (pdf)

$$\begin{aligned} p(y|\mu,\tau) &= \frac{1}{C(\mu,\tau)} \exp\left(-\frac{\tau}{2} d(\mu,y)^2\right), \\ C(\mu,\tau) &= \int_M \exp\left(-\frac{\tau}{2} d(\mu,y)^2\right) dy. \end{aligned} \tag{4.1}$$

We term this distribution a *Riemannian normal distribution*, and use the notation $y \sim N_M(\mu, \tau^{-1})$ to denote it. The parameter $\mu \in M$ acts as a location parameter on the manifold, and the parameter $\tau \in \mathbb{R}_+$ acts as a dispersion parameter, similar to the precision of a Gaussian. This distribution has the advantages that (1) it is applicable to any Riemannian manifold, (2) it reduces to a multivariate normal distribution (with isotropic covariance) when $M = \mathbb{R}^n$, and (3) much like the Euclidean normal distribution, maximum-likelihood estimation of parameters gives rise to least-squares methods (see [34] for details). We note that this noise model could be replaced with a different distribution, perhaps specific to the type of manifold or application, and the inference procedure presented in the next section could be modified accordingly.

The PPGA model for a random variable y on a smooth Riemannian manifold M is

$$y|x \sim N_M\left(\mathrm{Exp}(\mu, z), \tau^{-1}\right), z = W\Lambda x, \tag{4.2}$$

where $x \sim N(0,1)$ are again latent random variables in $\mathbb{R}^q$, μ here is a base point on M, W is a matrix with q columns of mutually orthogonal tangent vectors in $T_\mu M$, Λ is a $q \times q$ diagonal matrix of scale factors for the columns of W, and τ is a scale parameter for the noise. In this model, a linear combination of $W\Lambda$ and the latent variables x forms a new tangent vector $z \in T_\mu M$. Next, the exponential map shoots the base point μ by z

to generate the location parameter of a *Riemannian normal distribution*, from which the data point y is drawn. Note that in Euclidean space, the exponential map is an addition operation, $\text{Exp}(\mu, z) = \mu + z$. Thus, our model coincides with the standard PPCA model, when $M = \mathbb{R}^n$.

4.2.1 Inference

We develop a maximum likelihood procedure to estimate the parameters $\theta = (\mu, W, \Lambda, \tau)$ of the PPGA model defined in (4.2). Given observed data $y_i \in \{y_1, ..., y_N\}$ on M, with associated latent variable $x_i \in \mathbb{R}^q$, and $z_i = W\Lambda x_i$, we formulate an expectation maximization (EM) algorithm. Since the expectation step over the latent variables does not yield a closed-form solution, we develop a Hamiltonian Monte Carlo (HMC) method to sample x_i from the posterior $p(x|y;\theta)$, the log of which is given by

$$\log \prod_{i=1}^{N} p(x_i|y_i;\theta) \propto -N \log C - \sum_{i=1}^{N} \frac{\tau}{2} d\left(\text{Exp}(\mu, z_i), y_i\right)^2 - \frac{\|x_i\|^2}{2}, \tag{4.3}$$

and use this in a Monte Carlo Expectation Maximization (MCEM) scheme to estimate θ. The procedure contains two main steps:

4.2.1.1 E-step: HMC

For each x_i, we draw a sample of size S from the posterior distribution (4.3) using HMC with the current estimated parameters θ^k. Denote x_{ij} as the jth sample for x_i, the Monte Carlo approximation of the Q function is given by

$$Q(\theta|\theta^k) = E_{x_i|y_i;\theta^k}\left[\prod_{i=1}^{N} \log p(x_i|y_i;\theta^k)\right] \approx \frac{1}{S}\sum_{j=1}^{S}\sum_{i=1}^{N} \log p(x_{ij}|y_i;\theta^k). \tag{4.4}$$

In our HMC sampling procedure, the potential energy of the Hamiltonian $H(x_i, m) = U(x_i) + V(m)$ is defined as $U(x_i) = -\log p(x_i|y_i;\theta)$, and the kinetic energy $V(m)$ is a typical isotropic Gaussian distribution on a q-dimensional auxiliary momentum variable, m. This gives us a Hamiltonian system to integrate: $\frac{dx_i}{dt} = \frac{\partial H}{\partial m} = m$, and $\frac{dm}{dt} = -\frac{\partial H}{\partial x_i} = -\nabla_{x_i} U$. Due to the fact that x_i is a Euclidean variable, we use a standard "leap-frog" numerical integration scheme, which approximately conserves the Hamiltonian and results in high acceptance rates.

The computation of the gradient term $\nabla_{x_i} U(x_i)$ requires we compute $d_v \text{Exp}(p, v)$, i.e., the derivative operator (Jacobian matrix) of the exponential map with respect to the initial velocity v. The gradient with respect to each x_i is

$$\nabla_{x_i} U = x_i - \tau \Lambda W^T \{d_{z_i} \text{Exp}(\mu, z_i)^\dagger \text{Log}(\text{Exp}(\mu, z_i), y_i)\}, \tag{4.5}$$

where † represents the adjoint of a linear operator, i.e.,

$$\langle d_{z_i}\mathrm{Exp}(\mu, z_i)\hat{u}, \hat{v}\rangle = \langle \hat{u}, d_{z_i}\mathrm{Exp}(\mu, z_i)^\dagger \hat{v}\rangle.$$

4.2.1.2 M-step: Gradient Ascent

In this section, we derive the maximization step for updating the parameters $\theta = (\mu, W, \Lambda, \tau)$ by maximizing the HMC approximation of the Q function in (4.4). This turns out to be a gradient ascent scheme for all the parameters since there are no closed-form solutions.

The gradient of the Q function with respect to τ requires evaluation of the derivative of the normalizing constant in the *Riemannian normal distribution* (4.1). When M is a symmetric space, this constant does not depend on the mean parameter, μ, because the distribution is invariant to isometrics (see [34] for details). Thus, the normalizing constant can be written as

$$C(\tau) = \int_M \exp\left(-\frac{\tau}{2} d(\mu, y)^2\right) dy.$$

We can rewrite this integral in normal coordinates, which can be thought of as a polar coordinate system in the tangent space, $T_\mu M$. The radial coordinate is defined as $r = d(\mu, y)$, and the remaining $n-1$ coordinates are parametrized by a unit vector v, i.e., a point on the unit sphere $S^{n-1} \subset T_\mu M$. Thus, we have the change-of-variables, $\phi(rv) = \mathrm{Exp}(\mu, rv)$. Now the integral for the normalizing constant becomes

$$C(\tau) = \int_{S^{n-1}} \int_0^{R(v)} \exp\left(-\frac{\tau}{2} r^2\right) |\det(d\phi(rv))| dr\, dv, \tag{4.6}$$

where $R(v)$ is the maximum distance that $\phi(rv)$ is defined. Note that this formula is valid only if M is a complete manifold, which guarantees that normal coordinates are defined everywhere except possibly a set of measure zero on M.

The integral in (4.6) is difficult to compute for general manifolds, due to the presence of the determinant of the Jacobian of ϕ. However, for symmetric spaces this change-of-variables term has a simple form. If M is a symmetric space, there exists a orthonormal basis $u_1, \ldots, u_n$, with $u_1 = v$, such that

$$|\det(d\phi(rv))| = \prod_{k=2}^{n} \frac{1}{\sqrt{\kappa_k}} f_k(\sqrt{\kappa_k} r), \tag{4.7}$$

where $\kappa_k = K(u_1, u_k)$ denotes the sectional curvature, and f_k is defined as

$$f_k(x) = \begin{cases} \sin(x) & \text{if } \kappa_k > 0, \\ \sinh(x) & \text{if } \kappa_k < 0, \\ x & \text{if } \kappa_k = 0. \end{cases}$$

Notice that with this expression for the Jacobian determinant there is no longer a dependence on v inside the integral in (4.6). Also, if M is simply connected, then $R(v) = R$ does not depend on the direction v, and we can write the normalizing constant as

$$C(\tau) = A_{n-1} \int_0^R \exp\left(-\frac{\tau}{2} r^2\right) \prod_{k=2}^{n} \kappa_k^{-1/2} f_k(\sqrt{\kappa_k} r) dr,$$

where A_{n-1} is the surface area of the $n-1$ hypersphere, S^{n-1}. The remaining integral is one-dimensional, and can be quickly and accurately approximated by numerical integration. While this formula works only for simply connected symmetric spaces, other symmetric spaces could be handled by lifting to the universal cover, which is simply connected, or by restricting the definition of the Riemannian normal pdf in (4.1) to have support only up to the injectivity radius, i.e., $R = \min_v R(v)$.

The gradient term for estimating τ is

$$\nabla_\tau Q = \sum_{i=1}^{N} \sum_{j=1}^{S} \frac{1}{C(\tau)} A_{n-1} \int_0^R \frac{r^2}{2} \exp\left(-\frac{\tau}{2} r^2\right) \prod_{k=2}^{n} \kappa_k^{-1/2} f_k(\sqrt{\kappa_k} r) dr - \frac{1}{2} d(\mathrm{Exp}(\mu, z_{ij}), y_i)^2 dr.$$

From (4.3) and (4.4), the gradient term for updating μ is

$$\nabla_\mu Q = \frac{1}{S} \sum_{i=1}^{N} \sum_{j=1}^{S} \tau d_\mu \mathrm{Exp}(\mu, z_{ij})^\dagger \mathrm{Log}\left(\mathrm{Exp}(\mu, z_{ij}), y_i\right).$$

Here the derivative $d_\mu \mathrm{Exp}(\mu, v)$ is with respect to the base point, μ. Similar to (2.3), this derivative can be derived from a variation of geodesics: $c(s,t) = \mathrm{Exp}(\mathrm{Exp}(\mu, su), tv(s))$, where $v(s)$ comes from parallel translating v along the geodesic $\mathrm{Exp}(\mu, su)$. Again, the derivative of the exponential map is given by a Jacobi field satisfying $J_\mu(t) = dc/ds(0,t)$, and we have $d_\mu \mathrm{Exp}(\mu, v) = J_\mu(1)$.

For updating Λ, we take the derivative w.r.t. each ath diagonal element Λ^a as

$$\frac{\partial Q}{\partial \Lambda^a} = \frac{1}{S} \sum_{i=1}^{N} \sum_{j=1}^{S} \tau (W^a x_{ij}^a)^T \{ d_{z_{ij}} \mathrm{Exp}(\mu, z_{ij})^\dagger \mathrm{Log}(\mathrm{Exp}(\mu, z_{ij}), y_i) \},$$

where W^a denotes the ath column of W, and x_{ij}^a is the ath component of x_{ij}.

The gradient w.r.t. W is

$$\nabla_W Q = \frac{1}{S} \sum_{i=1}^{N} \sum_{j=1}^{S} \tau d_{z_{ij}} \mathrm{Exp}(\mu, z_{ij})^\dagger \mathrm{Log}(\mathrm{Exp}(\mu, z_{ij}), y_i) x_{ij}^T \Lambda. \tag{4.8}$$

To preserve the mutual orthogonality constraint on the columns of W, we represent W as a point on the Stiefel manifold $V_q(T_\mu M)$, i.e., the space of orthonormal q-frames in $T_\mu M$. We project the gradient in (4.8) onto the tangent space $T_W V_q(T_\mu M)$, and then update W

Algorithm 2: Monte Carlo Expectation Maximization for Probabilistic Principal Geodesic Analysis

Input: Dataset Y, reduced dimension q.
Initialize μ, W, Λ, σ.
repeat
  Sample X according to (4.5),
  Update μ, W, Λ, σ by gradient ascent in Section 3.2.2.
until convergence

by taking a small step along the geodesic in the projected gradient direction. For details on the geodesic computations for Stiefel manifolds, see [92].

The MCEM algorithm for PPGA is an iterative procedure for finding the subspace spanned by q principal components, shown in Algorithm 2. The computation time per iteration depends on the complexity of exponential map, log map, and Jacobi field which may vary for different manifold. Note the cost of the gradient ascent algorithm also linearly depends on the data size, dimensionality, and the number of samples drawn. An advantage of MCEM is that it can run in parallel for each data point.

4.3 Experiments

In this section, we demonstrate the effectiveness of PPGA and our ML estimation using both simulated data on the 2D sphere and a real corpus callosum data set. Before presenting the experiments of PPGA, we briefly review the necessary computations for the specific types of manifolds used, including the Riemannian exponential map, log map, and Jacobi fields.

4.3.1 Simulated Sphere Data

4.3.1.1 Sphere Geometry Overview

Let p be a point on an n-dimensional sphere embedded in $\mathbb{R}^{n+1}$, and let v be a tangent at p. The inner product between tangents at a base point p is the usual Euclidean inner product. The exponential map is given by a 2D rotation of p by an angle given by the norm of the tangent, i.e.,

$$\mathrm{Exp}(p, v) = \cos\theta \cdot p + \frac{\sin\theta}{\theta} \cdot v, \quad \theta = \|v\|. \tag{4.9}$$

The log map between two points p, q on the sphere can be computed by finding the initial velocity of the rotation between the two points. Let $\pi_p(q) = p \cdot \langle p, q \rangle$ denote the projection of the vector q onto p. Then,

$$\mathrm{Log}(p, q) = \frac{\theta \cdot (q - \pi_p(q))}{\|q - \pi_p(q)\|}, \quad \theta = \arccos(\langle p, q \rangle). \tag{4.10}$$

All sectional curvatures for S^n are equal to one. The adjoint derivatives of the exponential map are given by

$$d_p\text{Exp}(p,v)^\dagger w = \cos(\|v\|)w^\perp + w^\top, \qquad d_v\text{Exp}(p,v)^\dagger w = \frac{\sin(\|v\|)}{\|v\|}w^\perp + w^\top,$$

where $w^\perp, w^\top$ denote the components of w that are orthogonal and tangent to v, respectively. An illustration of geodesics and the Jacobi fields that give rise to the exponential map derivatives is shown in Figure 4.1.

4.3.1.2 Parameter Estimation on the Sphere

Using our generative model for PGA (4.2), we forward simulated a random sample of 100 data points on the unit sphere S^2, with known parameters $\theta = (\mu, W, \Lambda, \tau)$, shown in Table 4.1. Next, we ran our maximum likelihood estimation procedure to test whether we could recover those parameters. We initialized μ from a random uniform point on the sphere. We initialized W as a random Gaussian matrix, to which we then applied the Gram-Schmidt algorithm to ensure its columns were orthonormal. Figure 4.1 compares the ground truth principal geodesics and MLE principal geodesic analysis using our algorithm. A good overlap between the first principal geodesic shows that PPGA recovers the model parameters.

One advantage that our PPGA model has over the least-squares PGA formulation is that the mean point is estimated jointly with the principal geodesics. In the standard PGA algorithm, the mean is estimated first (using geodesic least-squares), and then the principal geodesics are estimated second. This does not make a difference in the Euclidean case (principal components must pass through the mean), but it does in the nonlinear case. We compared our model with PGA and standard PCA (in the Euclidean embedding space). The estimation error of principal geodesics turned out to be larger in PGA compared to our model. Furthermore, the standard PCA converges to an incorrect solution due to its inappropriate use of a Euclidean metric on Riemannian data. A comparison of the ground truth parameters and these methods is given in Table 4.1. Note that the noise precision τ is not a part of either the PGA or PCA models.

4.3.2 Shape Analysis of the Corpus Callosum

4.3.2.1 Shape Space Geometry

A configuration of k points in the 2D plane is considered as a complex k-vector, $z \in \mathbb{C}^k$. Removing translation, by requiring the centroid to be zero, projects this point to the linear complex subspace $V = \{z \in \mathbb{C}^k : \sum z_i = 0\}$, which is equivalent to the space $\mathbb{C}^{k-1}$. Next,

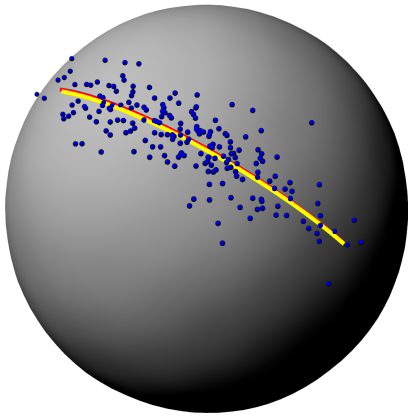


Figure 4.1: The principal geodesic of random generated data on unit sphere. Blue dots: random generated sphere dataset. Yellow line: ground truth principal geodesic. Red line: estimated principal geodesic using PPGA.

Table 4.1: Comparison between ground truth parameters for the simulated data and the MLE of PPGA, nonprobabilistic PGA, and standard PCA.

	μ	w	Λ	τ
Ground truth	$(-0.78, 0.48, -0.37)$	$(-0.59, -0.42, 0.68)$	0.40	100
PPGA	$(-0.78, 0.48, -0.40)$	$(-0.59, -0.43, 0.69)$	0.41	102
PGA	$(-0.79, 0.46, -0.41)$	$(-0.59, -0.38, 0.70)$	0.41	N/A
PCA	$(-0.70, 0.41, -0.46)$	$(-0.62, -0.37, 0.69)$	0.38	N/A

points in this subspace are deemed equivalent if they are a rotation and scaling of each other, which can be represented as multiplication by a complex number, $\rho e^{i\theta}$, where ρ is the scaling factor and θ is the rotation angle. The set of such equivalence classes forms the complex projective space, $\mathbb{C}P^{k-2}$.

We think of a centered shape $p \in V$ as representing the complex line $L_p = \{z \cdot p : z \in \mathbb{C}\backslash\{0\}\}$, i.e., L_p consists of all point configurations with the same shape as p. A tangent vector at $L_p \in V$ is a complex vector, $v \in V$, such that $\langle p, v\rangle = 0$. The exponential map is given by rotating (within V) the complex line L_p by the initial velocity v, that is,

$$\mathrm{Exp}(p, v) = \cos\theta \cdot p + \frac{\|p\| \sin\theta}{\theta} \cdot v, \quad \theta = \|v\|. \tag{4.11}$$

Likewise, the log map between two shapes $p, q \in V$ is given by finding the initial velocity of the rotation between the two complex lines L_p and L_q. Let $\pi_p(q) = p \cdot \langle p, q\rangle / \|p\|^2$ denote the projection of the vector q onto p. Then the log map is given by

$$\mathrm{Log}(p, q) = \frac{\theta \cdot (q - \pi_p(q))}{\|q - \pi_p(q)\|}, \quad \theta = \arccos \frac{|\langle p, q\rangle|}{\|p\|\|q\|}. \tag{4.12}$$

The sectional curvatures of $\mathbb{C}P^{k-2}$, $\kappa_i = K(u_i, v)$, used in (4.7), can be computed as follows. Let $u_1 = i \cdot v$, where we treat v as a complex vector and $i = \sqrt{-1}$. The remaining $u_2, \ldots, u_n$ can be chosen arbitrarily to construct an orthonormal frame with v and u_1, then we have $K(u_1, v) = 4$ and $K(u_i, v) = 1$ for $i > 1$. The adjoint derivatives of the exponential map are given by

$$d_p\mathrm{Exp}(p, v)^\dagger w = \cos(\|v\|)w_1^\perp + \cos(2\|v\|)w_2^\perp + w^\top,$$
$$d_v\mathrm{Exp}(p, v)^\dagger w = \frac{\sin(\|v\|)}{\|v\|}w_1^\perp + \frac{\sin(2\|v\|)}{2\|v\|} + w_2^\top,$$

where $w_1^\perp$ denotes the component of w parallel to u_1, i.e., $w_1^\perp = \langle w, u_1\rangle u_1$, $u_2^\top$ denotes the remaining orthogonal component of w, and $w^\top$ denotes the component tangent to v.

4.3.2.2 Shape Variability of Corpus Callosum Data

As a demonstration of PPGA on Kendall shape space, we applied it to corpus callosum shape data derived from the OASIS database (www.oasis-brains.org). The data consisted of magnetic resonance images (MRI) from 32 healthy adult subjects. The corpus callosum was segmented in a midsagittal slice using the ITK SNAP program (www.itksnap.org). An example of a segmented corpus callosum in an MRI is shown in Figure 4.2. The boundaries of these segmentations were sampled with 64 points using ShapeWorks (www.sci.utah.edu/software.html). This algorithm generates a sampling of a set of shape boundaries while enforcing correspondences between different point models within the population. Figure 4.2 displays the first two modes of corpus callosum shape variation, generated from points along the estimated principal geodesics: $\mathrm{Exp}(\mu, \alpha_i w_i)$, where $\alpha_i = -3\lambda_i, -1.5\lambda_i, 0, 1.5\lambda_i, 3\lambda_i$, for $i = 1, 2$.

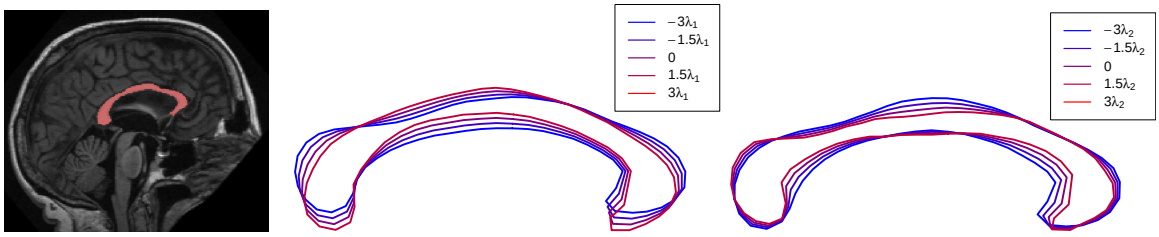


Figure 4.2: Left: example corpus callosum segmentation from an MRI slice. Middle to right: first and second PGA mode of shape variation with -3, -1.5, 1.5, and $3 \times \lambda$.

4.4 Automatic Data Dimensionality Reduction

Analogous to BPCA, this PPGA model can be extended to a Bayesian model of PGA by introducing a prior on the scale factor matrix Λ. This results in an automatic selection of model dimensionality. We add an automatic relevance determination (ARD) prior over each diagonal element Λ_i, with an associate precision hyperparameter γ_i, so that

$$p(\Lambda \mid \gamma) = \prod_{i=1}^{q} \frac{\gamma_i}{2\pi} \exp(-\frac{1}{2}\gamma_i \Lambda_i^2).$$

The value of γ_i is estimated iteratively as $1/\Lambda_i^2$. It enforces sparsity by driving the corresponding component Λ_i to zero, thus W_i will be effectively removed in the latent space. Notice that the columns of Λ define the principal subspace of standard PGA, therefore, inducing sparsity on Λ has the same effect as removing irrelevant dimensions in the principal subspace.

4.5 Conclusion

This chapter presented a latent variable model of PGA on Riemannian manifolds. We developed a Monte Carlo Expectation Maximization for maximum likelihood estimation of parameters that uses Hamiltonian Monte Carlo to integrate over the posterior distribution of latent variables. This work takes the first step to bring latent variable models to Riemannian manifolds. It opens up several possibilities for new factor analyses on Riemannian manifolds, including a rigorous formulation for mixture models of PGA and automatic dimensionality selection with a Bayesian formulation of PGA.

CHAPTER 5

BAYESIAN PRINCIPAL GEODESIC ANALYSIS OF DIFFEOMORPHIC SHAPE VARIABILITY

This chapter continues to present a generative Bayesian approach of principal geodesic analysis (PGA) for estimating the low-dimensional latent space of diffeomorphic shape variability in a population of images. We develop a latent variable model for PGA that provides a probabilistic framework for factor analysis in the space of infinite-dimensional diffeomorphisms. A sparsity prior in the model results in the automatic selection of the number of relevant dimensions by driving unnecessary principal geodesics to zero. To infer model parameters, including the image atlas, principal geodesic deformations, and the effective dimensionality, we introduce an expectation maximization (EM) algorithm.

5.1 Overview

Extracting low-dimensional, second-order statistics of anatomical shape variability is important to improve the statistical power and interpretability of further statistical analyses. The standard method for conducting dimensionality reduction and analyzing variability of Euclidean data is principal component analysis (PCA), which decomposes the data matrix into a linear combination of independent factors. Bishop [5] introduced a Bayesian model for PCA (BPCA) that automatically learns the dimension of the latent space from data by including a sparsity-inducing prior on each component of the factor matrix. These linear factor analysis models, nevertheless, are not directly applicable to nonlinear diffeomorphic transformations.

There exist several methods for dimensionality reduction and shape variability modeling on nonlinear manifolds. We reviewed principal geodesic analysis (PGA) proposed by Fletcher et al. [27] in Chapter 2. Based on this work, algorithms for exact solutions to PGA were developed in [87], [88]. In order to allow factor analysis on manifolds, we introduced

a probabilistic model for PGA (PPGA) in Chapter 4. In the setting of diffeomorphic image registration, Vaillant et al. [3] computed a tangent space PCA (TPCA) of the initial momenta from the atlas image. Later, Qiu et al. [4] used TPCA as an empirical shape prior in diffeomorphic surface matching. A Bayesian model of shape variability using diffeomorphic matching of currents is also formulated by Gori et al. [93]. Their model includes an estimation of a covariance matrix of the deformations, from which they then extracted PCA modes of shape variability. Even though these methods formulate the atlas and covariance estimation as probabilistic inference problems, the dimensionality reduction is done after the fact, i.e., as a singular value decomposition of the covariance as a second stage after the estimation step.

We propose instead to treat the dimensionality reduction step as a probabilistic inference problem on discrete images, in a model called Bayesian principal geodesic analysis (BPGA), which jointly estimates the image atlas and principal geodesic modes of variation. Our model goes beyond the PPGA algorithm by introducing automatic dimensionality reduction, as well as extending from finite-dimensional manifolds to the infinite-dimensional case of diffeomorphic image registration. This Bayesian formulation has two advantages. First, it explicitly optimizes the fit of the principal modes to the data intrinsically in the space of diffeomorphisms, which results in better fits to the data. Second, by formulating dimensionality reduction as a Bayesian model with a sparsity prior, we can also infer the inherent dimensionality directly from the data.

In this chapter, we incorporate a stronger sparsity prior, based on the adaptive sparsity method of Figueiredo [94] that avoids the need for hyperparameters, and provide in-depth derivations of the statistical model and inference procedure. We also mention the relationship of our work to manifold learning approaches and dimensionality reduction methods in [95], [96]. Unlike the nonparametric manifold learning methods, the Bayesian approach we present here is parametric and fully *generative*. The shape deformation of individuals is explicitly encoded in the model, and can be reconstructed directly in a compact space of principal modes of deformations. We show experimental results of principal geodesics and parameters estimated from both 2D synthetic data and 3D OASIS brain MRI data. To validate the advantages of our model, we reconstruct images from our estimation and compare the reconstruction errors with TPCA of diffeomorphisms and LPCA based on image intensity. Our results indicate that intrinsic modeling of the principal geodesics, estimated jointly with the image atlas, provides a better description of brain image data than computing PCA in the tangent space after atlas estimation.

5.2 Probability Model

We formulate the random initial velocity for the kth individual as $v^k = Wx^k$, where W is a matrix with q columns of principal initial velocities, and $x^k \in \mathbb{R}^q$ is a latent variable that lies in a low-dimensional space, with

$$p(x^k \,|\, W) \propto \exp\left(-\frac{1}{2}\left\|Wx^k\right\|_V^2\right). \tag{5.1}$$

Compared to BPCA, the difference of this latent variable prior is incorporating W as a conditional probability, which guarantees smoothness of the geodesic shooting path. Notice that we shift from the momenta space in [54] to a nicely smooth velocity space, which gains more stable computations.

Our noise model is based on the assumption of i.i.d. Gaussian at each image voxel, much like [4], [71], [58]. This can be varied under different conditions, for instance, spatially dependent models for highly correlated noise data. In this chapter, we will focus on the commonly used and simple Gaussian noise model, with the likelihood given by

$$p(J^k \,|\, I, \sigma, x^k) = \frac{1}{(2\pi)^{M/2}\sigma^M} \exp\left(-\frac{\left\|I \circ (\phi^k)^{-1} - J^k\right\|_{L^2}^2}{2\sigma^2}\right), \tag{5.2}$$

where M is the number of voxels, and the norm inside the exponent is the $L^2(\Omega, \mathbb{R})$ norm. Note that for a continuous image domain, $\Omega = \mathbb{R}^d/\mathbb{Z}^d$, this is not a well-defined probability distribution due to its infinite measure in the Hilbert space $L^2(\Omega, \mathbb{R})$ on images. Therefore, we consider the input images as well as diffeomorphisms to be defined on a finite discretized grid.

The prior on W is a sparsity prior that suppresses the small principal initial velocity to zero. This prior is analogous to the hierarchical sparsity prior proposed by [94], with the difference that we use the natural Hilbert space norm for the velocity. The prior is based on Laplacian distribution, a widely used and exploited way to achieve sparse estimation. It presses the irrelevant or redundant components exactly to zero. As first introduced by [97], the Laplace distribution is equivalent to the marginal distribution of a hierarchical-Bayes model: a Gaussian prior with zero mean and exponentially distributed variances. Let i denote the ith principal component of W. We define each component W_i as a random variable with the hierarchical model distribution

$$\begin{aligned} p(W_i \,|\, \tau_i) &\sim N(0, \tau_i), \\ p(\tau_i \,|\, \gamma_i) &\sim \mathrm{Exp}(\frac{\gamma_i}{2}), \end{aligned}$$

After integrating out τ_i, we have the marginalized distribution as

$$p(W_i \,|\, \gamma_i) = \int_0^\infty p(W_i \,|\, \tau_i) p(\tau_i \,|\, \gamma_i) d\tau_i = \frac{\sqrt{\gamma_i}}{2} \exp\left(-\sqrt{\gamma_i}\, \|W_i\|_1\right),$$

which is a Laplacian distribution with scale parameter $\gamma_i/2$. The degree of sparsity is controlled by the hyperparameter γ_i on the l_1 penalty. However, the sparsity parameter is specified in an ad hoc manner. [94] proposed an effective model to remove γ_i by adopting a Jeffreys' noninformative hyperprior as $p(\tau_i) \sim 1/\tau_i$. This has the advantages that (1) the improper hyperprior is scale-invariant, (2) the model is parameter-free. Using this hierarchical sparsity prior on the columns of W for the automatic dimensionality selection, we formulate the problem as

$$\begin{aligned} p(W, x \,|\, \tau) &\propto \exp\left(-\frac{1}{2}\sum_{k=1}^{N} \left\|Wx^k\right\|_V^2 - \sum_{i=1}^{q} \frac{\|W_i\|_V^2}{2\tau_i}\right), \\ p(\tau) &\propto \frac{1}{\tau}, \end{aligned} \tag{5.3}$$

where $x = [x^1, ..., x^k]$, $\tau = [\tau_1, ..., \tau_q]$. We will later integrate out the latent variable τ using expectation maximization.

We can express our model for the kth subject using the graphical representation shown in Figure 5.1.

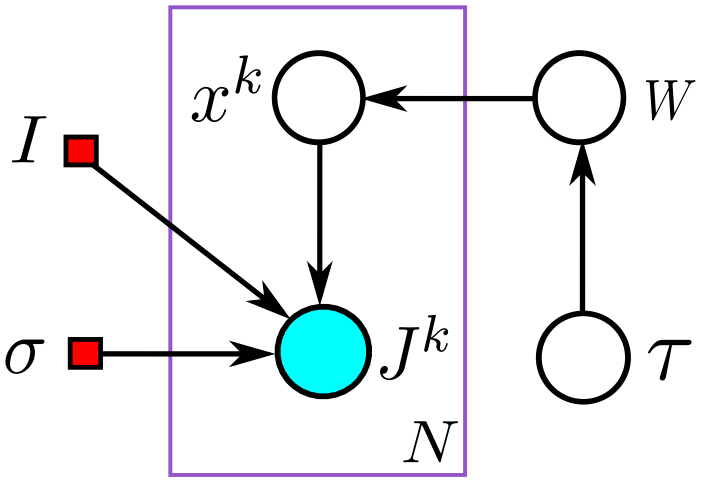


Figure 5.1: Graphical representation of BPGA for the kth subject J^k.

5.3 Inference

We use MAP estimation to determine the model parameters $\theta = \{I, \sigma\}$. After defining the likelihood (5.2) and prior (5.3) in the previous section, we now arrive at the joint posterior for BPGA as

$$\prod_{k=1}^{N} p\left(W, x, \tau \,|\, J^k; \theta\right) \;\propto\; \left[\prod_{k=1}^{N} p(J^k \,|\, x^k, \theta)\, p(x^k \,|\, W)\right] p(W \,|\tau)\, p(\tau). \tag{5.4}$$

In order to treat the W, x^k and τ as latent random variables with the log posterior given by (5.4), we would ideally integrate out the latent variables, which are intractable in closed form for W, x^k. Instead, we develop an expectation maximization algorithm to compute a closed-form solution to integrate out τ first, and then use a mode approximation for W, x^k to the posterior distribution. It contains two alternating steps:

- **E-step** Using the current estimate of the parameters $\hat{\theta}$, we compute the expectation Q of the complete log-posterior of (5.4) with respect to the latent variables τ as

$$\begin{aligned} Q(W, x^k, \theta \,|\, \hat{\theta}, \hat{W}) \;\propto\; & -\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2 - \frac{MN}{2} \log \sigma \\ & - \frac{1}{2} \sum_{k=1}^{N} \left\| W x^k \right\|_V^2 - \sum_{i=1}^{q} \frac{\|W_i\|_V^2}{2\|\hat{W}_i\|_V^2}. \end{aligned} \tag{5.5}$$

  Note that we use the same approach to integrate out τ in [94]. Details are in Appendix A.

- **M-step: Gradient Ascent for** W, x^k We introduce a gradient ascent scheme to estimate W, x^k, and $\theta = (I, \sigma)$ simultaneously. We need to compute the gradient with respect to the initial velocity v^k of the diffeomorphic image matching problem in (2.18), and then apply the chain rule to obtain the gradient term w.r.t. W and x^k. Following the optimal control theory approach in [1], we add Lagrange multipliers to constrain the kth diffeomorphism ϕ_t^k to be a geodesic path, which is done by introducing time-dependent adjoint variables, $\hat{I}_t^k$, $\hat{m}_t^k$, $\hat{v}_t^k$ for transported image I_t^k, momentum m_t^k, and velocity v_t^k, respectively. To make the calculation simple to read, we drop the notation t and denote $\partial_t f$ as $\dot{f}$ for any function f. We then write the augmented energy

$$\begin{aligned} \tilde{Q}(W, x^k, \theta \,|\, \hat{\theta}, \hat{W}) = \; & Q + \sum_{k=1}^{N} \int_0^1 \Big[\langle \hat{v}^k, \dot{v}^k + K \mathrm{ad}_{v^k}^* m^k \rangle_{L^2} \\ & + \langle \hat{I}^k, \dot{I}^k + \nabla I^k \cdot v^k \rangle_{L^2} + \langle \hat{m}^k, m^k - L v^k \rangle_{L^2} \Big] dt, \end{aligned} \tag{5.6}$$

where Q is the expectation function from (5.5), and the other terms correspond to Lagrange multipliers enforcing a) the geodesic constraint, which comes from the EPDiff equation (2.13), b) the image transport equation, $\dot{I}^k = -\nabla I^k \cdot v^k$, and c) the constraint, $m^k = Lv^k$, respectively.

Dropping out the terms that are not related to W, x^k and I_0 in (5.6), we have

$$\begin{aligned}\tilde{Q}(W, x^k, \theta \,|\, \hat{\theta}, \hat{W}) \propto & -\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left\|I\circ(\phi^k)^{-1} - J^k\right\|_{L^2}^2 - \frac{1}{2}\sum_{k=1}^{N}\left\|Wx^k\right\|_V^2 \\ & -\sum_{i=1}^{q}\frac{\|W_i\|_V^2}{2\|\hat{W}_i\|_V^2} + \sum_{k=1}^{N}\int_0^1 \Big[\langle \hat{v}^k, \dot{v}^k + K\mathrm{ad}^*_{v^k}m^k\rangle_{L^2} \\ & + \langle \hat{I}^k, \dot{I}^k + \nabla I^k \cdot v^k\rangle_{L^2} + \langle \hat{m}^k, m^k - Lv^k\rangle_{L^2}\Big]dt. \end{aligned} \tag{5.7}$$

The gradient of $\tilde{Q}$ with respect to the kth initial velocity is $\nabla_{v^k}\tilde{Q} = v^k - K\hat{v}^k$ (details are in Appendix A). Applying the chain rule, the gradient term of (5.7) for updating W is

$$\nabla_W \tilde{Q} = -\sum_{k=1}^{N}(v^k - K\hat{v}^k)(x^k)^T - W\Lambda,$$

where Λ is a diagonal matrix with diagonal element $\frac{1}{\|\hat{W}_i\|_V^2}$. The gradient with respect to x^k is

$$\nabla_{x^k}\tilde{Q} = -W^T(v^k - K\hat{v}^k).$$

- **Closed-form solution for θ** We now derive the maximization for updating the parameters θ. This turns out to be a closed-form update for the atlas I, noise variance σ^2. For updating I and σ, we set the derivative of the expectation with respect to I, σ to zero (see Appendix A). The solution for I, σ^2 gives an update

$$I = \frac{\sum_{k=1}^{N} J^k \circ \phi^k |D\phi^k|}{\sum_{k=1}^{N}|D\phi^k|}, \quad \sigma^2 = \frac{1}{MN}\sum_{k=1}^{N}\left\|I\circ(\phi^k)^{-1} - J^k\right\|_{L^2}^2.$$

5.4 Results

We demonstrate the effectiveness of our proposed model and MAP estimation routine using both 2D synthetic data and real 3D MRI brain data.

5.4.1 Synthetic Data

Because we have a generative model, we can forward simulate a random sample of images from a distribution with known parameters $\theta = (I, \sigma)$. We tested if we can recover those parameters using our BPGA inference procedure. We simulated a 2D synthetic dataset with 40 subjects starting from an atlas image, I, of a binary circle with resolution 100×100. We then generated random samples of W with two principal modes and x^k from the prior distribution, $p(W, x^k \,|\, \tau)$, defined in (5.3), setting $\alpha = 0.2$ for the Laplacian operator L. To generate a deformed circle image, we shot the initial velocity constructed by Wx^k, and transformed the atlas by the resulting diffeomorphisms. Finally, we added i.i.d. Gaussian noise according to our likelihood model (5.2). We used a standard deviation of $\sigma = 0.05$, which corresponds to a SNR of 20 (which is more noise than typical structural MRI). Figure 5.2 compares the ground truth atlas I and principal geodesics with our estimation. In addition, our estimation of the noise variance $\sigma = 0.051$ is also close to the ground truth $\sigma = 0.05$. It shows that our method recovers the model parameters fairly well. Figure 5.3 demonstrates the shape variation of our synthetic dataset from the atlas by $a_i = -3, -1.5, 0, 1.5, 3$.

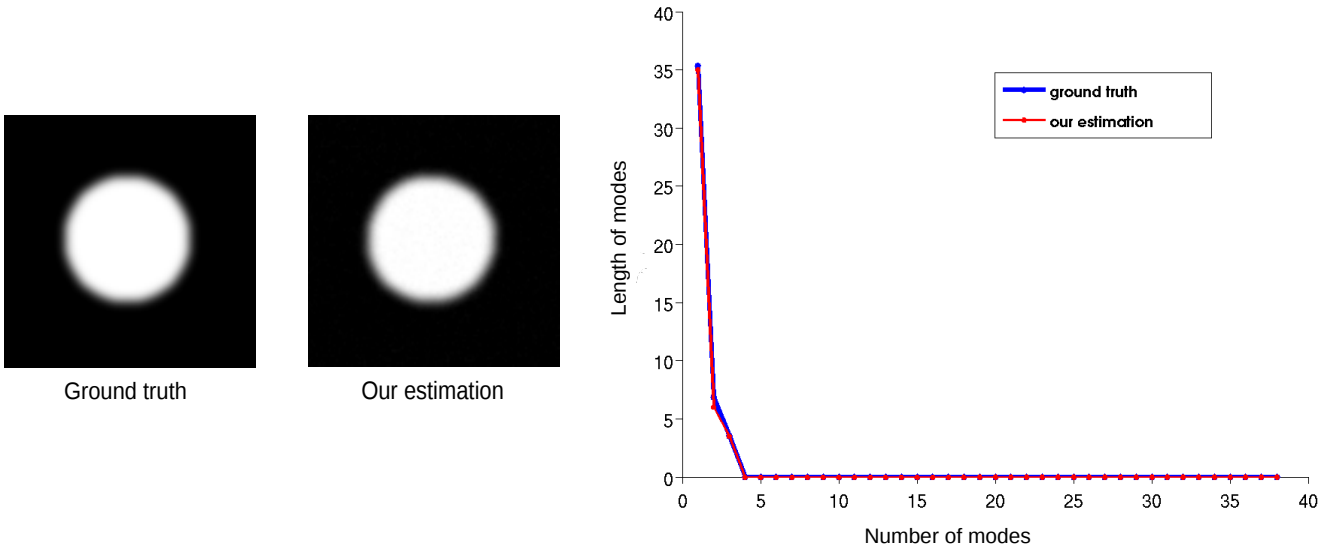


Figure 5.2: Left to right: ground truth of atlas I; our estimation of atlas; ground truth of the length of all principal geodesics and our estimation.

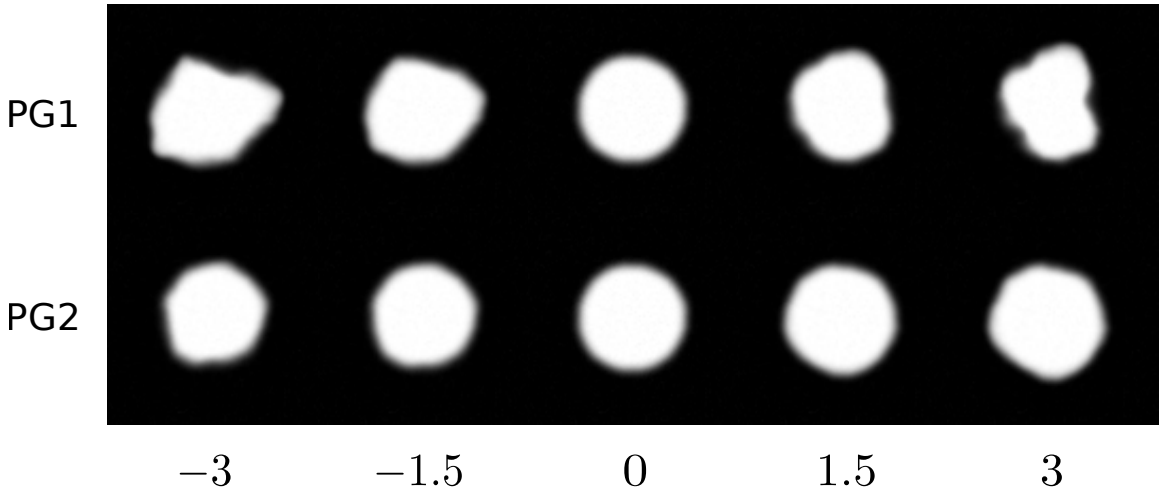


Figure 5.3: Top to bottom: shooting atlas by the first and second principal geodesics. Left to right: BPGA model of image variation evaluated at $a_i = -3, -1.5, 0, 1.5, 3$.

5.4.2 OASIS Brain Dataset

To demonstrate the effectiveness of our proposed model and MAP estimation, we applied our BPGA model to a set of brain magnetic resonance images (MRI) from the 3D OASIS brain database. The data consist of MRIs from 130 subjects between the age of 60 to 95. The MRIs have a resolution of $128 \times 128 \times 128$ with an image spacing of $1.0 \times 1.0 \times 1.0$ mm^3 and are skull-stripped, intensity normalized, and co-registered with rigid transforms. To set the parameters in L operator, we did the initial step of estimating $\alpha = 0.1$ using the procedure in [58]. We used 15 time-steps in geodesic shooting and initialize the template I as the average of image intensities, with W as the matrix of principal components from TPCA.

The proposed BPGA model automatically determined that the latent dimensionality of the data was 15. Figure 5.4 displays the automatically estimated modes, $i = 1, 2$, of the brain MRI variation. We forward shoot the constructed atlas, I, by the estimated principal momentum $a_i W_i$ along the geodesics. For the purpose of visualization, we demonstrate the brain variation from the atlas by $a_i = -3, -1.5, 0, 1.5, 3$. We also show the log determinant of Jacobians at $a_i = 3$, with red representing regions of expansion and blue representing regions of contraction. The first mode of variation clearly shows that ventricle size change is a dominant source of variability in brain shape. Our algorithm also jointly estimated the image noise standard deviation parameter as $\sigma = 0.04$.

We validated the ability of our BPGA model to compactly represent the space of brain variations by testing how well it can reconstruct unseen images. After estimating

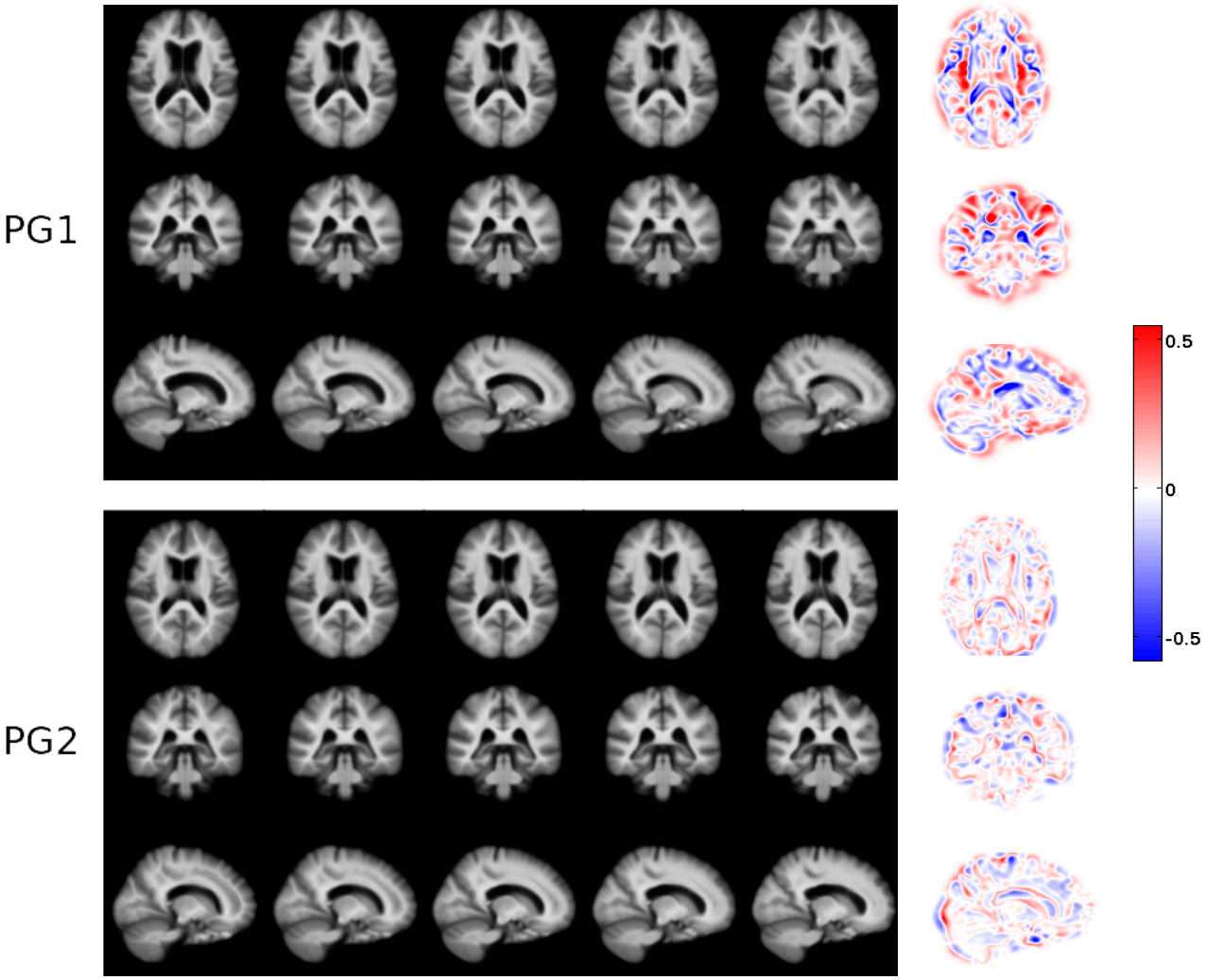


Figure 5.4: Top to bottom: axial, coronal and sagittal views of shooting the atlas by the first and second principal modes. Left to right: BPGA model of image variation evaluated at $a_i = -3, -1.5, 0, 1.5, 3$, and log determinant of Jacobians at $a_i = 3$.

the principal initial velocity and parameters from the training subjects above, we used these estimates to reconstruct another 20 testing subjects from the same OASIS database that were not included in the training. We then measured the discrepancy between the reconstructed images and the unobserved testing images. Note that our reconstruction used only the first fifteen principal modes, which were automatically selected by our algorithm.

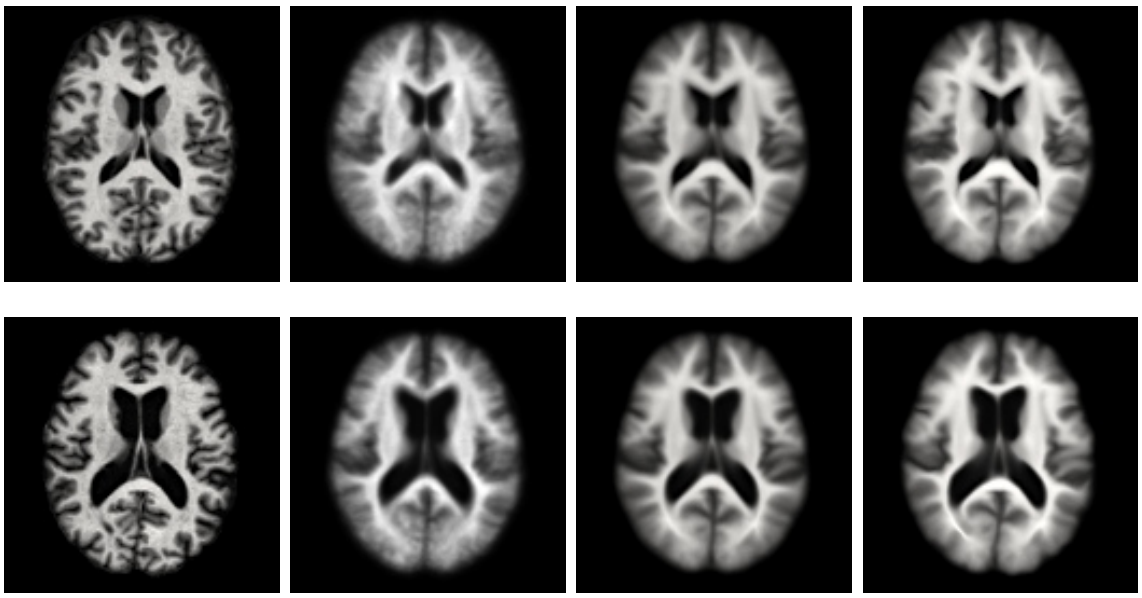
We use the first fifteen dimensions to compare our model with LPCA and TPCA. Table 5.1 shows the comparison of the reconstruction accuracy as measured by the average and standard deviation of the mean squared error (MSE). The table indicates that our model outperforms both LPCA and TPCA in the diffeomorphic setting.

Table 5.1: Comparison of mean squared reconstruction error between LPCA, TPCA, and BPGA models. Average and standard deviation over 20 test images.

	LPCA	TPCA	BPGA
Average MSE	4.2×10^{-2}	3.4×10^{-2}	2.8×10^{-2}
Std of MSE	1.25×10^{-2}	4.8×10^{-3}	4.2×10^{-3}

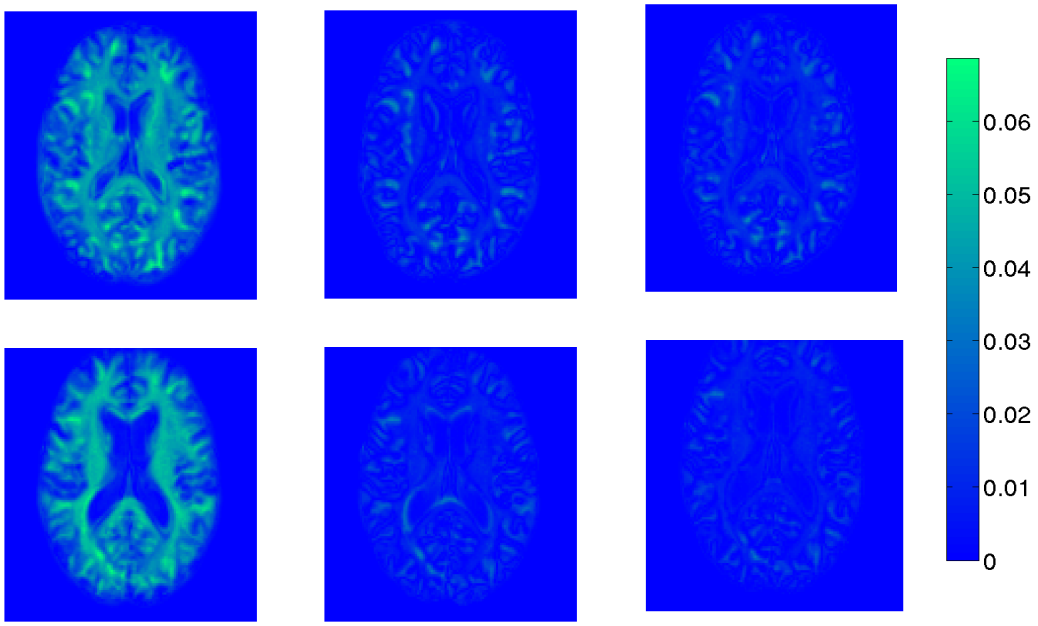
Examples of the reconstructed images are shown in Figure 5.5. The model BPGA recovers a better shape of brain structures than the other two models. Next, Figure 5.6 displays the maps of the absolute value of reconstruction error by LPCA, TPCA, and our model BPGA. It clearly demonstrates that BPGA has less reconstruction error of the unknown dataset than both LPCA and TPCA.

We also display the reconstruction error with increasing number of principal modes from 5 to 40. Notice that our model BPGA automatically only keeps the selected 15 modes but eliminates all others. Therefore, we do not have reconstruction error with BPGA when the number of modes goes beyond 15. Figure 5.7 shows that TPCA requires approximately 32 principal modes, more than twice as many as our model does, to achieve the same level of reconstruction accuracy. The model LPCA cannot match the BPGA reconstruction accuracy with even 40 principal modes. This reflects that our model BPGA gains a more compact representation than TPCA and LPCA.



(a) Observed (b) LPCA (c) TPCA (d) BPGA

Figure 5.5: Left to right: original data, reconstruction by LPCA, TPCA, and BPGA.



(a) LPCA (b) TPCA (c) BPGA

Figure 5.6: Left to right: absolute value of reconstruction error map by LPCA, TPCA, and BPGA.

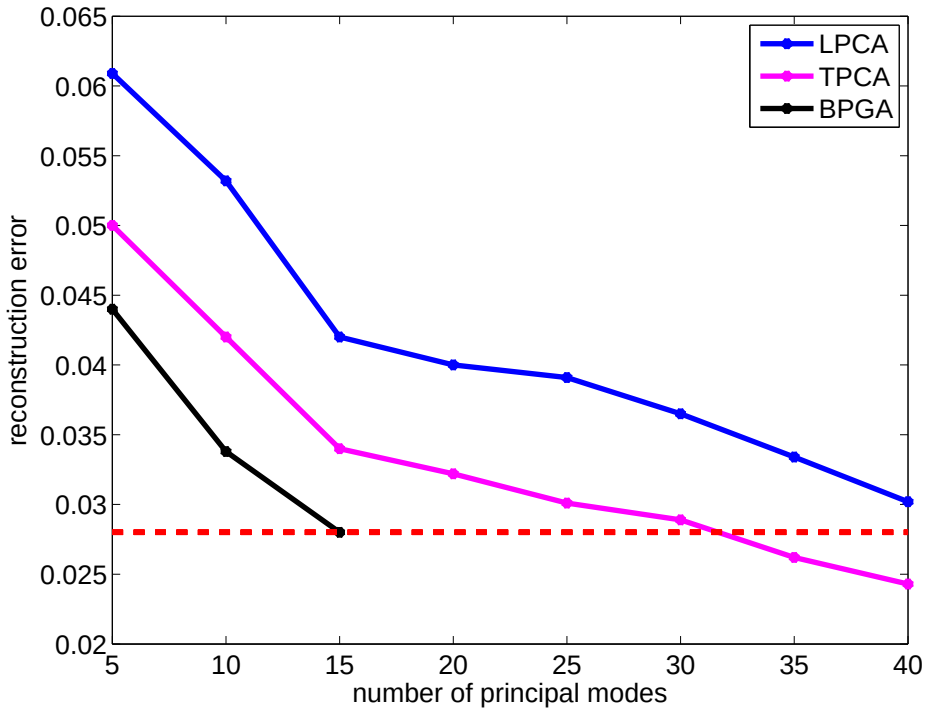


Figure 5.7: Averaged mean squared reconstruction error with a different number of principal modes by LPCA, TPCA, and BPGA over 20 test images.

5.5 Conclusion and Future Work

This chapter presented a generative Bayesian model of principal geodesic analysis in diffeomorphic image registration. Our method is the first probabilistic model for automatic dimensionality reduction for diffeomorphisms. We developed an inference strategy based on MAP to estimate parameters, including the noise variance and image atlas, simultaneously. The estimated low-dimensional latent variables provide a compact representation of the anatomical variability in a large image database, and they can be used for further statistical analysis of anatomical shape in clinical studies. Reducing the dimensionality to the inherent modes of shape variability has the potential to improve hypothesis testing, classification, and mixture models.

There are several avenues for future work to build upon our BPGA model. In this chapter, we precomputed the regularization parameter using the simple atlas building model in [58]. Since different parameters can lead to different principal modes, atlas, etc., ideally we would estimate the regularization parameter simultaneously with all other parameters. Doing this would require a more computationally expensive approach that integrates out the latent x variables, rather than the mode approximation used here. Such an approach has been done for PPGA on finite-dimensional manifolds [53]. This would be related to several other approaches that integrate out deformations in image atlas building. For instance, [65] proposed a fully generative Bayesian model of small elastic deformation in which the latent image transformations are marginalized from the distribution. Markov chain Monte Carlo (MCMC) methods for sampling elastic deformations in Bayesian atlas models have been introduced by [67], [69], and [98]. Furthermore, [70] inferred the regularization parameter from a hierarchical Bayesian model, although their work was in the elastic deformation setting as well. [58] were the first to develop a truly Bayesian model for diffeomorphic atlas building and regularization parameter estimation by integrating out latent random diffeomorphisms.

In addition, much like the Euclidean BPCA model [5], we did not enforce that the principal modes be orthogonal. This can be achieved by optimization in the Stiefel manifold of orthonormal frames, as is done in [53]. However, the high-dimensionality of velocity fields makes this a difficult problem to implement directly.

CHAPTER 6

LOW DIMENSIONAL LIE ALGEBRAS FOR GEODESIC SHOOTING

The Bayesian inference of diffeomorphisms comes with a high computational cost because of marginalizing the high-dimensional diffeomorphisms on dense spatial grids. To solve this problem, this chapter introduces an algorithm FLASH, which defines a novel definition of low-dimensional Lie algebras in the space of bandlimited velocity fields. A key ingredient is that we compute all the geodesic evolution equations and adjoint Jacobi field equations needed for gradient descent methods entirely in these low-dimensional Lie algebras. Another important factor of FLASH is a reduced version of adjoint Jacobi field equations that gives fast convergence. We not only speed up the current geodesic shooting algorithm dramatically, but also require much less memory than state-of-the-art methods. Previous work has modeled the continuous variational problem of diffeomorphisms, and then discretized them to solve on a computer. A discrete representation being a Lie algebra itself has never been considered before. This work is the first to present a finite-dimensional Lie algebra as the discrete approximation to the tangent space of infinite-dimensional diffeomorphisms. To demonstrate the effectiveness of our model, we test a pairwise registration on both 2D synthetic data and 3D brain images, and compare its convergence, run-time, and memory consumption with leading LDDMM methods. The experimental results show that FLASH is not only faster than state-of-the-art LDDMM algorithms, but also converges to better solutions, i.e., lower values of the registration objective function. The work of this chapter is also presented and published in [99].

6.1 Overview

LDDMM has been applied and becomes an indispensable tool in many fields of medical image analysis, for instance, atlas building, shape variation quantification, atlas-based image segmentation, etc. However, LDDMM comes with a huge cost for dense spatial grids. The integration over the velocity field at each time point results in a high computational cost

and large memory footprint, especially in large imaging studies. In addition, a gradient descent optimization of the time-dependent velocity field is used to compute the geodesic path in [52]. This first requires a gradient term with expensive numerical solutions to partial differential equations. Furthermore, the optimization might get stuck in a local minimum due to the high-dimensional dense grid. Convergence can be very slow as well. To develop a more efficient optimization scheme for LDDMM, Vialard et al. [1] introduced a geodesic shooting algorithm where only the initial velocity at time point zero was estimated. Ashburner [6] showed that a Gauss-Newton implementation gained a faster convergence. All these approaches were based on the fact that a geodesic is uniquely determined by its initial velocity via the geodesic evolution equations. In this case, the initial velocity is sufficient to parameterize the geodesic. Even though geodesic shooting avoids storing the entire time-varying velocity field at each iteration, the geodesic shooting and backward integration of adjoint equations needed for gradient evaluation are still computationally expensive on a high-dimensional dense grid.

One approach to alleviate the computational requirement of calculating diffeomorphisms is stationary velocity fields, introduced by Arsigny et al. [100]. The major contribution of stationary velocity fields is defining another parameterization of diffeomorphisms through velocity fields that remain constant in time. With this representation, the flow of diffeomorphisms is generated by solving stationary ordinary differential equations (ODEs). This one-parameter subgroup parameterization of diffeomorphisms reduces the computational cost and memory demands in the original LDDMM framework. Based on this, Ashburner [101] later proposed a fast diffeomorphic image registration, called DARTEL. The solution was estimated efficiently by composing successive diffeomorphisms over time using a scaling and squaring approach, which was originally proposed by Arsigny et al. [100]. A full multigrid strategy was added in the optimization for achieving fast convergence and avoiding local minima. Stationary velocity fields save more time and memory than LDDMM, but they do not provide distance metrics on the space of diffeomorphisms. Greedy algorithms have also been studied for speeding up diffeomorphic registration. They iteratively apply gradient updates to a single deformation field instead of a full time-dependent flow at each iteration. The most popular algorithms in this category are the original diffeomorphic image registration [7] and the diffeomorphic demons algorithm presented by Vercauteren et al. [8]. Greedy methods are much faster and more memory efficient in exploring a large solution space, but they do not minimize a global variational problem. They also lack the definition of geodesics parameterized by the initial velocity field, and thus do not provide a distance

metric between images. Such a parameterization paves the way for statistical models of diffeomorphisms in a linear vector space, for instance, principal component analysis [3] and regression [102].

Previous work has focused on the continuous theorem of infinite-dimensional Lie algebras of diffeomorphisms. Researchers later approximated it on a finite-dimensional discrete grid for real implementations on the computer, while in this chapter, we are the first to formulate a real discrete representation of diffeomorphisms directly via low-dimensional Lie algebras in the space of bandlimited velocity fields. This new concept of low-dimensional Lie algebras has two main contributions. First, it allows us to compute all the geodesic evolution equations and adjoint Jacobi field equations needed for gradient descent methods entirely in a low-dimensional space. Second, it preserves the distance metric of LDDMM. This chapter incorporates run-time complexity to further discuss where exactly FLASH gains speed and adds convergence analysis. We also mention the relationship of our work to discrete parameterization of diffeomorphisms by a finite set of control points [103]. Instead of estimating the geometric position of each control point, we prespecify an effective dimension of space to work with directly, which saves more computational cost. To demonstrate the effectiveness of our model, we test a pairwise registration on both 2D synthetic data and 3D brain images and compare its convergence, run-time, and memory consumption with the leading LDDMM method. The experimental results show that FLASH is not only faster than the state-of-the-art LDDMM algorithm, but also converges to better solutions, i.e., lower values of the registration objective function.

6.2 Low-Dimensional Lie Algebras

In this section, we introduce a low-dimensional Lie algebra with its corresponding Lie bracket that gives a discrete representation of vector fields.

The key observation is that the velocity fields in the geodesic evolution equation, also known as the EPDiff equation (2.13), stay in a low frequency domain. The K operator as the last calculation in (2.13) is a low-pass filter, and as such, suppresses high frequency components in the velocity fields (see Figure 6.1). This means that the velocity fields are already bandlimited to a certain maximum frequency. However, the previous implementation of geodesic shooting, using full-dimensional velocity fields, wastes effort computing the high frequency components, which just end up being forced to zero by K. We instead propose to develop a low-dimensional discretization of velocity fields as bandlimited signals in the Fourier domain.

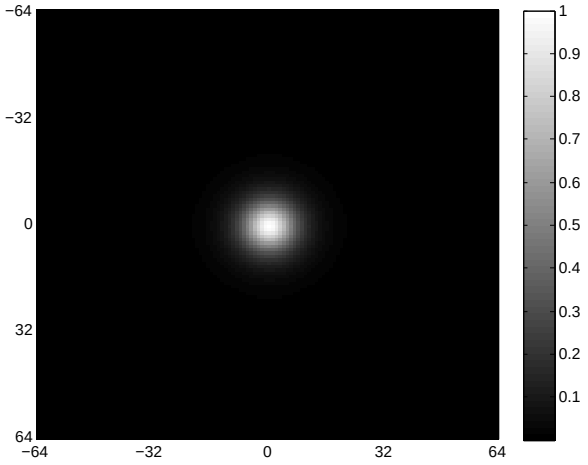


Figure 6.1: Fourier coefficients of the discretized K operator on a 128×128 grid, with parameters $\alpha = 3$, $c = 3$.

6.2.1 Space of Bandlimited Velocity Fields and Metrics

Let $\tilde{V}$ denote the space of bandlimited velocity fields on Ω, with frequency bounds $N_1, N_2, \ldots, N_d$ in each of the dimensions of Ω. Any element $\tilde{v} \in \tilde{V}$ is a multidimensional array: $\tilde{v}_{k_1,k_2,\ldots,k_d} \in \mathbb{C}^d$, where $k_i \in 0, \ldots, N_i - 1$ is the frequency index along the ith axis. Note that to ensure $\tilde{v}$ represents a real-valued vector field in the spatial domain, we have the constraint that $\tilde{v}_{k_1,\ldots,k_d} = \tilde{v}^*_{N_1-k_1,\ldots,N_d-k_d}$, where $*$ denotes the complex conjugate.

There is a natural projection mapping, $\nu : V \to \tilde{V}$, of V into the space $\tilde{V}$ of complex vector fields, given by the Fourier series expansion:

$$\nu(v)(k) = \int_{-\infty}^{\infty} v(x) e^{-j2\pi xk}\, dx, \tag{6.1}$$

where $k = (k_1, \ldots, k_d)$, and $x = (x_1, \ldots, x_d)$ is the spatial index.

Vice versa, the inclusion mapping $\iota : \tilde{V} \to V$, of $\tilde{V}$ into the space V is given by:

$$\iota(\tilde{v})(x) = \sum_{0 \leqslant k < N} \tilde{v}(k) e^{j2\pi kx}\, dk. \tag{6.2}$$

Note that we use a multi-index notation for a simple expression of the original formulation, which is

$$\sum_{0 \leqslant k < N} \tilde{v}(k) e^{j2\pi kx}\, dk = \sum_{k_1=0}^{N_1-1}, \ldots, \sum_{k_d=0}^{N_d-1} \tilde{v}_{k_1,\ldots,k_d} e^{j2\pi k_1 x_1}, \ldots, e^{j2\pi k_d x_d}\, dk_1, \ldots, dk_d.$$

The metric at identity on $\tilde{V}$ is given by the discretized version of the $\tilde{L}$ operator as

$$\langle \tilde{v}, \tilde{w} \rangle_{\tilde{V}} = \int_{\Omega} \langle \tilde{L}\tilde{v}(x), \tilde{w}(x) \rangle dx.$$

The Fourier transformation of $L = (-\alpha\Delta + I)^c$ is a diagonal operator. Discretizing this operator by only keeping the frequencies up to our bandlimits, N_i, we get a diagonal

matrix $\tilde{L}$. Analogous to the L operator, this $\tilde{L} : \tilde{V} \to \tilde{V}^*$ maps a tangent vector in the Fourier domain to its dual momentum vector $\tilde{m}$. For a 3D grid, the coefficient $\tilde{L}_{k_1k_2k_3}$ of this operator in the Fourier domain is

$$\tilde{L}_{k_1k_2k_3} = \left[-2\alpha\left(\cos\frac{2\pi k_1}{N_1} + \cos\frac{2\pi k_2}{N_2} + \cos\frac{2\pi k_3}{N_3}\right) + 7\right]^c.$$

The Fourier coefficient of the K operator is $\tilde{K}_{k_1k_2k_3} = \tilde{L}^{-1}_{k_1k_2k_3}$.

To define a right-invariant metric at any other point ϕ besides identity, we map the velocity fields back to a full spatial domain by (6.1) first, and then pull back by the right composition with ϕ^{-1} to identity. This is because the Lie group that associates with our discrete low-dimensional Lie algebra is not a diffeomorphism group. Despite this disadvantage, our method still benefits hugely from the fact that a large portion of expensive computations in diffeomorphic image registration can be done in the low-dimensional Lie algebra.

A time sequence of bandlimited velocity fields in $\tilde{V}$ can consequently generate a flow of diffeomorphisms, $t \mapsto \phi_t \in \mathrm{Diff}^\infty(\Omega)$, in the following way. Using the inclusion mapping $\iota : \tilde{V} \to V$ defined in (6.2), we can generate the diffeomorphic flow as

$$\frac{d\phi_t(x)}{dt} = \iota(\tilde{v}_t) \circ \phi_t(x), \qquad x \in \Omega. \tag{6.3}$$

6.2.2 Low-Dimensional Lie Bracket

We now define a discrete Lie bracket that is analogous to the continuous operator in (2.14).

Definition 9 *For any two vector fields $\tilde{v}, \tilde{w} \in \tilde{V}$, a low-dimensional Lie bracket in the discrete Fourier domain is*

$$[\tilde{v},\, \tilde{w}] = (\tilde{D}\tilde{v}) * \tilde{w} - (\tilde{D}\tilde{w}) * \tilde{v}, \tag{6.4}$$

where $\tilde{D}\tilde{v}$ is the central difference Jacobian matrix of a discrete vector field.

This Jacobian matrix can be computed as a tensor product $\tilde{D}\tilde{v} = \eta \otimes \tilde{v}$ in the discrete Fourier domain, with $\eta \in \tilde{V}$ given by

$$\eta_{k_1,k_2,\ldots,k_d} = (j\sin(2\pi k_1), \ldots, j\sin(2\pi k_d)).$$

Due to the fact that the pointwise multiplication of two vector fields in the spatial domain corresponds to convolution in the Fourier domain, we can easily decompose the multiplication of a square matrix and vector field in the Fourier domain as a single convolution

for each row of the matrix. For notation simplicity, we denote all matrix-vector field and vector-vector field convolution as $*$ (see (B.1) in Appendix B for an explicit formulation of $*$). Note that because a convolution between two bandlimited signals does not preserve the bandlimit, we must follow a convolution operation by truncation back to the bandlimits, N_i, in each dimension to guarantee this Lie bracket is closed.

Next, we prove that this discrete operation satisfies the axioms to be a valid Lie algebra on $\tilde{V}$.

Theorem 1 *The vector space $\tilde{V}$, when equipped with the bracket operation (6.4), is a discrete low-dimensional Lie algebra. That is to say, $\forall\, \tilde{u},\, \tilde{v},\, \tilde{w} \in \tilde{V}$ and $a, b \in \mathbb{R}$, the following properties are satisfied:*

(a) Linearity: $[a\tilde{u} + b\tilde{v},\, \tilde{w}] = a\,[\tilde{u},\, \tilde{w}] + b\,[\tilde{v},\, \tilde{w}]$,

(b) Anticommutativity: $[\tilde{u},\, \tilde{v}] = -[\tilde{v},\, \tilde{u}]$,

(c) Jacobi identity: $[\tilde{u},\, [\tilde{v},\, \tilde{w}]] + [\tilde{w},\, [\tilde{u},\, \tilde{v}]] + [\tilde{v},\, [\tilde{w},\, \tilde{u}]] = 0$.

Proof: Linearity and anticommutativity are immediate. We have

(a) Linearity:

$$\begin{aligned}
[a\tilde{u} + b\tilde{v},\, \tilde{w}] &= \tilde{D}(a\tilde{u} + b\tilde{v}) * \tilde{w} - \tilde{D}\tilde{w} * (a\tilde{u} + b\tilde{v}) \\
&= a(\tilde{D}\tilde{u} * \tilde{w} - \tilde{D}\tilde{w} * \tilde{u}) + b(\tilde{D}\tilde{v} * \tilde{w} - \tilde{D}\tilde{w} * \tilde{v}) \\
&= a[\tilde{u},\, \tilde{w}] + b[\tilde{v},\, \tilde{w}],
\end{aligned}$$

(b) Anticommutativity:

$$[\tilde{u},\, \tilde{v}] = \tilde{D}\tilde{u} * \tilde{v} - \tilde{D}\tilde{v} * \tilde{u} = -\left(\tilde{D}\tilde{v} * \tilde{u} - \tilde{D}\tilde{u} * \tilde{v}\right) = -[\tilde{v},\, \tilde{u}].$$

(c) Jacobi identity: The proof of the Jacobi identity follows closely that of the continuous case. First, note that the iterated central difference operator results in a third-order tensor:

$$\tilde{D}^2\tilde{u} = \tilde{D}\tilde{D}\tilde{u} = \eta \otimes \eta \otimes \tilde{u}.$$

Much like the Hessian tensor of a vector-valued function in the continuous case, the discrete Hessian is also symmetric with respect to contraction with a pair of vectors. That is,

$$\tilde{D}^2\tilde{u} * \tilde{v} * \tilde{w} = \tilde{D}^2\tilde{u} * \tilde{w} * \tilde{v},$$

where the convolution between $\tilde{D}^2\tilde{u}$ and $\tilde{v}$ is now analogous to the pointwise contraction of a third-order tensor field with a vector field in the spatial domain.

Next, we note that the product rule of differentiation also carries over to the discrete Fourier representation, and we have the identity

$$\tilde{D}(\tilde{D}\tilde{u} * \tilde{v}) = \tilde{D}^2\tilde{u} * \tilde{v} + \tilde{D}\tilde{u} * \tilde{D}\tilde{v},$$

where the second convolution operator is analogous to pointwise matrix field multiplication in the spatial domain. We then have

$$\begin{aligned}
[\tilde{u}, [\tilde{v}, \tilde{w}]] &= [\tilde{u}, \tilde{D}\tilde{v} * \tilde{w} - \tilde{D}\tilde{w} * \tilde{v}] \\
&= \tilde{D}\tilde{u} * (\tilde{D}\tilde{v} * \tilde{w} - \tilde{D}\tilde{w} * \tilde{v}) - \tilde{D}(\tilde{D}\tilde{v} * \tilde{w} - \tilde{D}\tilde{w} * \tilde{v}) * \tilde{u} \\
&= \tilde{D}\tilde{u} * \tilde{D}\tilde{v} * \tilde{w} - \tilde{D}\tilde{u} * \tilde{D}\tilde{w} * \tilde{v} - \tilde{D}^2\tilde{v} * \tilde{w} * \tilde{u} - \tilde{D}\tilde{v} * \tilde{D}\tilde{w} * \tilde{u} \\
&\quad + \tilde{D}^2\tilde{w} * \tilde{v} * \tilde{u} + \tilde{D}\tilde{w} * \tilde{D}\tilde{v} * \tilde{u} \qquad (6.5)
\end{aligned}$$

Similarly, we rewrite the other two terms as

$$\begin{aligned}
[\tilde{w}, [\tilde{u}, \tilde{v}]] &= \tilde{D}\tilde{w} * \tilde{D}\tilde{u} * \tilde{v} - \tilde{D}\tilde{w} * \tilde{D}\tilde{v} * \tilde{u} - \tilde{D}^2\tilde{u} * \tilde{v} * \tilde{w} - \tilde{D}\tilde{u} * \tilde{D}\tilde{v} * \tilde{w} \\
&\quad + \tilde{D}^2\tilde{v} * \tilde{u} * \tilde{w} + \tilde{D}\tilde{v} * \tilde{D}\tilde{u} * \tilde{w} \qquad (6.6)
\end{aligned}$$

$$\begin{aligned}
[\tilde{v}, [\tilde{w}, \tilde{u}]] &= \tilde{D}\tilde{v} * \tilde{D}\tilde{w} * \tilde{u} - \tilde{D}\tilde{v} * \tilde{D}\tilde{u} * \tilde{w} - \tilde{D}^2\tilde{w} * \tilde{u} * \tilde{v} - \tilde{D}\tilde{w} * \tilde{D}\tilde{u} * \tilde{v} \\
&\quad + \tilde{D}^2\tilde{u} * \tilde{w} * \tilde{v} + \tilde{D}\tilde{u} * \tilde{D}\tilde{w} * \tilde{v} \qquad (6.7)
\end{aligned}$$

Finally, by combining the equations (6.5), (6.6), (6.7), and using the symmetric rule above, we obtain

$$[\tilde{u}, [\tilde{v}, \tilde{w}]] + [\tilde{w}, [\tilde{u}, \tilde{v}]] + [\tilde{v}, [\tilde{w}, \tilde{u}]] = 0.$$

6.2.3 EPDiff Equation in Low-Dimensional Lie Algebras

Analogous to the EPDiff equation (2.13), we define a geodesic evolution equation in the discrete Fourier domain as

$$\frac{\partial \tilde{v}}{\partial t} = -\mathrm{ad}^{\dagger}_{\tilde{v}}\tilde{v} = -\tilde{K}\mathrm{ad}^{*}_{\tilde{v}}\tilde{m}, \qquad (6.8)$$

where the operator $\mathrm{ad}^* : \tilde{V}^* \to \tilde{V}^*$ is the dual of the negative low-dimensional Lie bracket of vector fields in the Fourier space, and its discrete formulation is

$$\mathrm{ad}^{*}_{\tilde{v}}\tilde{m} = (\tilde{D}\tilde{v})^T \star \tilde{m} + \tilde{\Gamma}(\tilde{m} \otimes \tilde{v}), \qquad (6.9)$$

where $\star$ denotes a truncated autocorrelation, and $\tilde{\Gamma}$ is a discrete divergence of a vector field $\tilde{v}$. It is computed as the sum of the pointwise multiplication $\tilde{\Gamma}\tilde{v} = \sum_{0 \leqslant k < N} \tilde{v}\tilde{\eta}^T$, where η is

the Fourier coefficient of the central differential operator for each dimension. Details for deriving this ad* operator are in Appendix B.

Plugging (6.9) back into the geodesic evolution equation (6.8), we have

$$\frac{\partial \tilde{v}}{\partial t} = -\mathrm{ad}^{\dagger}_{\tilde{v}}\tilde{v} = -\tilde{K}\left[(\tilde{D}\tilde{v})^T \star \tilde{m} + \tilde{\Gamma}(\tilde{m} \otimes \tilde{v})\right]. \tag{6.10}$$

6.3 Estimation of Diffeomorphic Image Registration

Now we are ready to present a *geodesic shooting* algorithm for diffeomorphic image registration using our low-dimensional Lie algebras. This is a gradient descent algorithm on the initial velocity $\tilde{v}_0 \in \tilde{V}$. Geodesic shooting of $\tilde{v}_0$ proceeds entirely in the reduced low-dimensional Lie algebra, producing a time-varying velocity, $t \mapsto \tilde{v}_t \in \tilde{V}$. A flow of diffeomorphic transformations is then generated by (6.3). This leads to a modification for the energy function (2.16) for LDDMM, where we now parameterize diffeomorphisms by the low-dimensional velocity $\tilde{v}_0$:

$$E(\tilde{v}_0) = \frac{1}{2\sigma^2}\|I_0 \circ \phi_1^{-1} - I_1\|^2 + (\tilde{L}\tilde{v}_0, \tilde{v}_0). \tag{6.11}$$

Before describing the details of our diffeomorphic image matching algorithm, we first provide an outline of the general steps. Beginning with the initialization $\tilde{v}_0 = 0$, the gradient descent algorithm to minimize the energy (6.11) proceeds by iterating the following:

1. **Forward shooting of $\tilde{v}_0$:** forward integrate the geodesic evolution equations (6.10) on $\tilde{V}$ to generate $\tilde{v}_{t_k}$ at discrete time points $t_1 = 0, t_2, \ldots, t_T = 1$.

2. **Compute the inverse diffeomorphism ϕ_1^{-1}:** compute the inverse diffeomorphism, ϕ_1^{-1}, by integrating the negative velocity field backward in time.

3. **Compute initial conditions for backward integration:** compute the initial condition for the adjoint variable $\hat{I}(1) = \frac{1}{\sigma^2}(I(1) - I_1)$ at $t = 1$.

4. **Bring gradient to $t = 0$ by reduced adjoint Jacobi field:** integrate the reduced adjoint Jacobi field equations in $\tilde{V}$ to get the gradient update $\nabla_{\tilde{v}_0} E$.

Note that steps 2 and 3 are computed at the full resolution of the input images in the spatial domain. However, steps 1 and 4 are computed entirely in the low-dimensional space $\tilde{V}$, resulting in greatly reduced computation time and memory requirements. We now provide details for the computations in each of these steps.

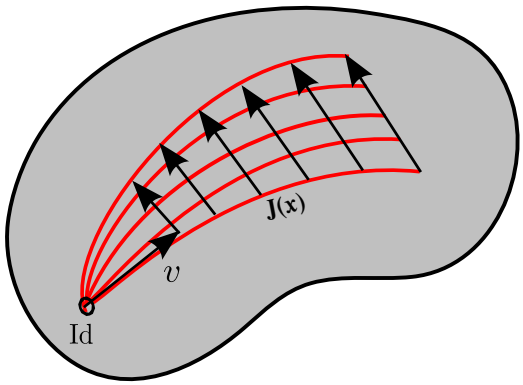


Figure 6.2: Jacobi fields

To generate ϕ_t^{-1}, we integrate the negative velocity fields backward under the left invariant metric (see details in [104]) as

$$\frac{d\phi_t^{-1}(x)}{dt} = -D\phi_t^{-1}(x) \cdot \iota(\tilde{v}_t), \qquad x \in \Omega.$$

As derived in [52], the gradient $\nabla_{\tilde{v}_1} E$ at time point $t = 1$ is computed after using the inclusion mapping $\iota : V \to \tilde{V}$ in (6.1) as

$$\nabla_{\tilde{v}_1} E = \nu \left(-K(\frac{1}{\sigma^2}(I_0 \circ \phi_1^{-1} - I_1) \cdot \nabla(I_0 \circ \phi_1^{-1})) \right). \tag{6.12}$$

We next introduce reduced adjoint Jacobi fields in the bandlimited velocity space to integrate the gradient term (6.12) at $t = 1$ backward to the initial point $t = 0$.

6.3.1 Reduced Adjoint Jacobi Fields in Bandlimited Velocity Space

Consider a variation of geodesics $\gamma : (-\epsilon, \epsilon) \times [0, 1] \to \text{Diff}(\Omega)$, with initial conditions $\gamma(0, t) = \phi_t$ and $\gamma(p, 0) = \text{Id}$, which is the diffeomorphic transformation at identity. Such a variation corresponds to a variation of the initial velocity $(d/dt)\gamma(p, 0) = v_0 + p\delta v_0$. The variation $\gamma(p, t)$ produces a "fan" of geodesics, illustrated in Figure 6.2. Taking the derivative of this variation results in a Jacobi field: $J_v(t) = d\gamma/dp(0, t)$.

In this chapter, we use a reduced version of adjoint Jacobi fields from [20], which is also used by [21]. A big advantage of using reduced adjoint Jacobi fields is that we can also decouple images from velocity fields in the backward integration. This is different from the vector momenta LDDMM [57], where images are jointly integrated backwards with velocity fields. We show that using this reduced adjoint Jacobi field results in an even better convergence rate than vector momenta LDDMM.

Under the right invariant metric of diffeomorphisms, we define a vector field $U(t) \in \tilde{V}$ as a right trivialized reduced Jacobi field $U(t) = \nu \left(J_v(t)\phi_t^{-1} \right)$, and a variation of the right

trivialized reduced velocity $\tilde{v}$ is $\delta\tilde{v}$. These variables satisfy the following reduced Jacobi equations:

$$\frac{d}{dt}\begin{pmatrix} U \\ \delta\tilde{v} \end{pmatrix} = \begin{pmatrix} \mathrm{ad}_{\tilde{v}} & I \\ 0 & \mathrm{sym}_{\tilde{v}} \end{pmatrix}\begin{pmatrix} U \\ \delta\tilde{v} \end{pmatrix}, \tag{6.13}$$

where $\mathrm{sym}_{\tilde{v}}\delta\hat{v} = -\mathrm{ad}^{\dagger}_{\tilde{v}}\delta\hat{v} - \mathrm{ad}^{\dagger}_{\delta\hat{v}}\tilde{v}$.

To transport the gradient term $\nabla_{\tilde{v}_1}E$ backward to the space of initial velocity fields, we use reduced Jacobi fields that are simply computed by the adjoint of the reduced Jacobi equations in (6.13). This results in another ordinary differential equation (ODE) as

$$\frac{d}{dt}\begin{pmatrix} \hat{U} \\ \delta\hat{v} \end{pmatrix} = \begin{pmatrix} -\mathrm{ad}^{\dagger}_{\tilde{v}} & 0 \\ -I & -\mathrm{sym}^{\dagger}_{\tilde{v}} \end{pmatrix}\begin{pmatrix} \hat{U} \\ \delta\hat{v} \end{pmatrix}, \tag{6.14}$$

where $\hat{U}, \delta\hat{v} \in \tilde{V}$ are introduced adjoint variables, and $\mathrm{sym}^{\dagger}_{\tilde{v}}\delta\hat{v} = -\mathrm{ad}_{\tilde{v}}\delta\hat{v} + \mathrm{ad}^{\dagger}_{\delta\hat{v}}\tilde{v}$. For more details on the derivation of the reduced adjoint Jacobi field equations, see [20].

Given the initial conditions $\hat{U}(1) = 0$ and $\delta\hat{v}(1) = \nabla_{\tilde{v}_1}E$, we obtain the transported gradient $\delta\hat{v}(0)$ by integrating the adjoint ODE (6.14) backward in time to $t = 0$.

Finally, we arrive at the gradient of E w.r.t. $\tilde{v}_0$ as

$$\nabla_{\tilde{v}_0}E = \tilde{v}_0 + \delta\hat{v}(0).$$

6.4 Complexity Analysis

Before presenting the theoretical complexity of FLASH and vector momenta LDDMM [57], we first review the gradient descent steps for vector momenta LDDMM. The gradients w.r.t. initial momenta are computed by adding Lagrange multipliers: a) the geodesic constraint that comes from the EPDiff equation (2.13), b) the image transport equation along time t, and c) the enforcement of $m = Lv$ to constrain the diffeomorphism $\phi(t)$ to be a geodesic path. This is done by optimizing an augmented energy function with introduced time-dependent adjoint variables for the initial momenta, deformed image, and initial velocity, respectively. The detailed steps are:

1. **Forward shooting:** forward integrate the geodesic evolution equations (2.13) on a dense grid to generate v_{t_k} at discrete time points $t_1 = 0, t_2, \ldots, t_T = 1$.

2. **Compute the diffeomorphism ϕ_t:** compute and store the diffeomorphism ϕ_t by integrating (2.12) forward in time.

3. **Compute the inverse diffeomorphism ϕ_t^{-1}:** compute the inverse diffeomorphism ϕ_t^{-1} by fixed-point iteration.

4. **Compute initial conditions for backward integration:** compute the gradient, $\nabla_{v_1} E$, of the energy (2.16) at $t = 1$.

5. **Bring gradient to $t = 0$:** integrate the adjoint equations in a high-dimensional space V to update $\nabla_{v_0} E$.

Table 6.1 demonstrates the comparison of computational complexity and memory requirement between FLASH with the number of voxels q and vector momenta LDDMM for pairwise 3D image registration with the number of voxels Q. It shows that the computations of forward and backward shooting of velocity fields are dramatically lower in FLASH on a downsampled grid than vector momenta LDDMM on a full dense grid. Furthermore, our model obtains the inverse of diffeomorphism ϕ_t^{-1} directly from integrating the negative bandlimted velocity fields backwards under the left invariant metric of diffeomorphisms, without wasting the computational cost of getting ϕ_t first. Even though the complexity of computing ϕ^{-1} and initial conditions for backward integration in vector momenta LDDMM shows better than our algorithm, the exact run-time is worse in practice due to the unspecified constant factors implied in the big-O notations. To address this issue, we compared the empirical running time of the most expensive operation fast Fourier transformation in FLASH and linear interpolation in vector momenta LDDMM with different scales of image dimension. Table 6.2 displays that the linear interpolation costs more than twice the number of a Fourier transform. We also measure the exact computational time and memory consumption for both our algorithm and vector momenta LDDMM through testing results on real 3D images in the following section.

Table 6.1: Comparison with vector momenta LDDMM for the computational complexity and memory requirement.

	Complexity		Memory	
	FLASH	LDDMM	FLASH	LDDMM
Forward shooting	$O(Tdq \log dq)$	$O(TdQ \log dQ)$	$O(dq)$	$O(dQ)$
Compute ϕ	0	$O(TdQ)$	0	$O(dQ)$
Compute ϕ^{-1}	$O\left(TdQ \log dQ\right)$	$O\left(TdQ\right)$	$O(dQ)$	$O(dQ)$
Initial conditions	$O\left(dQ \log dQ\right)$	$O\left(dQ\right)$	$O(dQ)$	$O(dQ)$
Backward integration	$O(Tdq \log dq)$	$O(TdQ \log dQ)$	$O(dq)$	$O(dQ)$

Table 6.2: Exact run-time comparison between Fourier transform and grid interpolation at different scale of dimension N.

	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
Fourier transform (s)	$7 \times e^{-4}$	$6.7 \times e^{-3}$	0.07	0.76	7.03
Interpolation (s)	$1.7 \times e^{-3}$	$1.66 \times e^{-2}$	0.15	1.98	18.7

6.5 Results

To demonstrate the effectiveness of FLASH, we use both 2D synthetic data and real 3D OASIS MRI brain data. All the experiments set $\alpha = 3.0, c = 3.0, \sigma = 0.03$ with $T = 10$ time-steps for geodesic shooting.

6.5.1 Synthetic Data

We tested 2D pairwise image registration of sine waves with different frequencies $f = 2, 4, \ldots, 16$ using our model FLASH and compared with vector momenta LDDMM. We set the truncated dimension $N = 32$ according to the number of nonzero Fourier coefficients of K operator shown in Figure 6.1. The image matching from FLASH with $N = 32, 128$ and vector momenta LDDMM in Figure 6.3 demonstrates that (1) our model FLASH is able to capture the same amount of information in a much lower-dimensional bandlimited space as in a full-dimensional space, (2) FLASH gains better image matching than vector momenta LDDMM as the frequency of sine waves increases. However, both methods fail the registration when the frequency goes up to $f = 16$ (see Figure 6.4). Notice that because of the constraint of smoothness on the transformations, sharp corners in the target images get smoothed in both vector momenta LDDMM and FLASH estimation.

6.5.2 3D Brain Image Registration

In this section, we first tested FLASH for pairwise image registration at different levels of truncated dimension $N = 4, 8, 16, 32, 64, 128$. The MRIs have resolution $128 \times 128 \times 128$ and are skull-stripped, intensity normalized, and co-registered with rigid transforms. We compared the total energy formulated in (2.16), time consumption, and memory requirement of our model versus the open source implementation of vector momenta LDDMM [57] (https://bitbucket.org/scicompanat/vectormomentum). For peer-to-peer comparison, we use the same integration method and (α, c, σ, T) parameters for both models.

Figure 6.5 displays the comparison of total energy, time, and memory at different levels of truncated dimensions. It indicates that our model FLASH gains a better image matching result but with much less time and memory. We see that our method achieves a lower overall

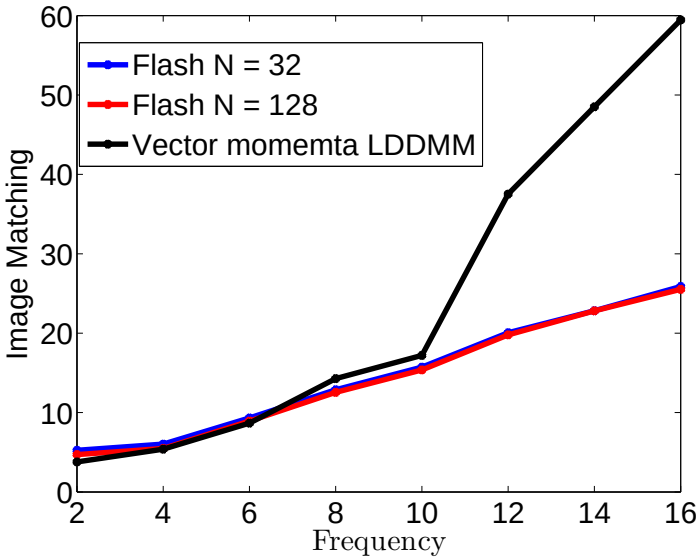


Figure 6.3: Comparison of image matching on sine wave images at different frequencies $f = 2, 4, \ldots, 16$ between FLASH with truncated dimension $N = 32$, full dimension $N = 128$, and vector momenta LDDMM.

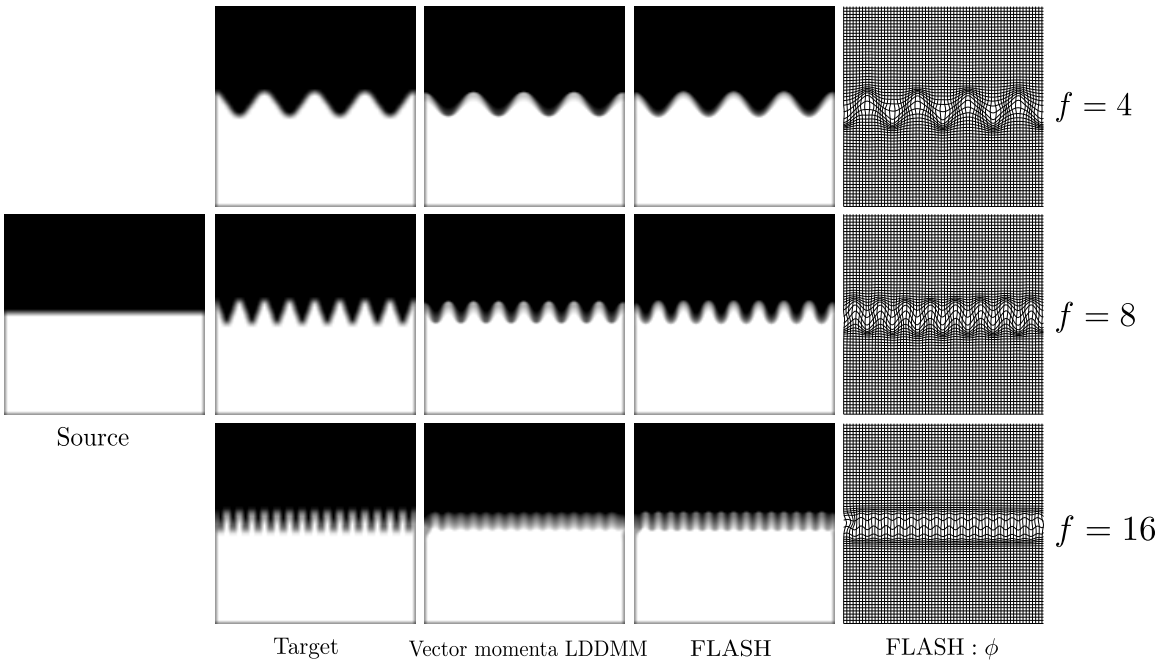
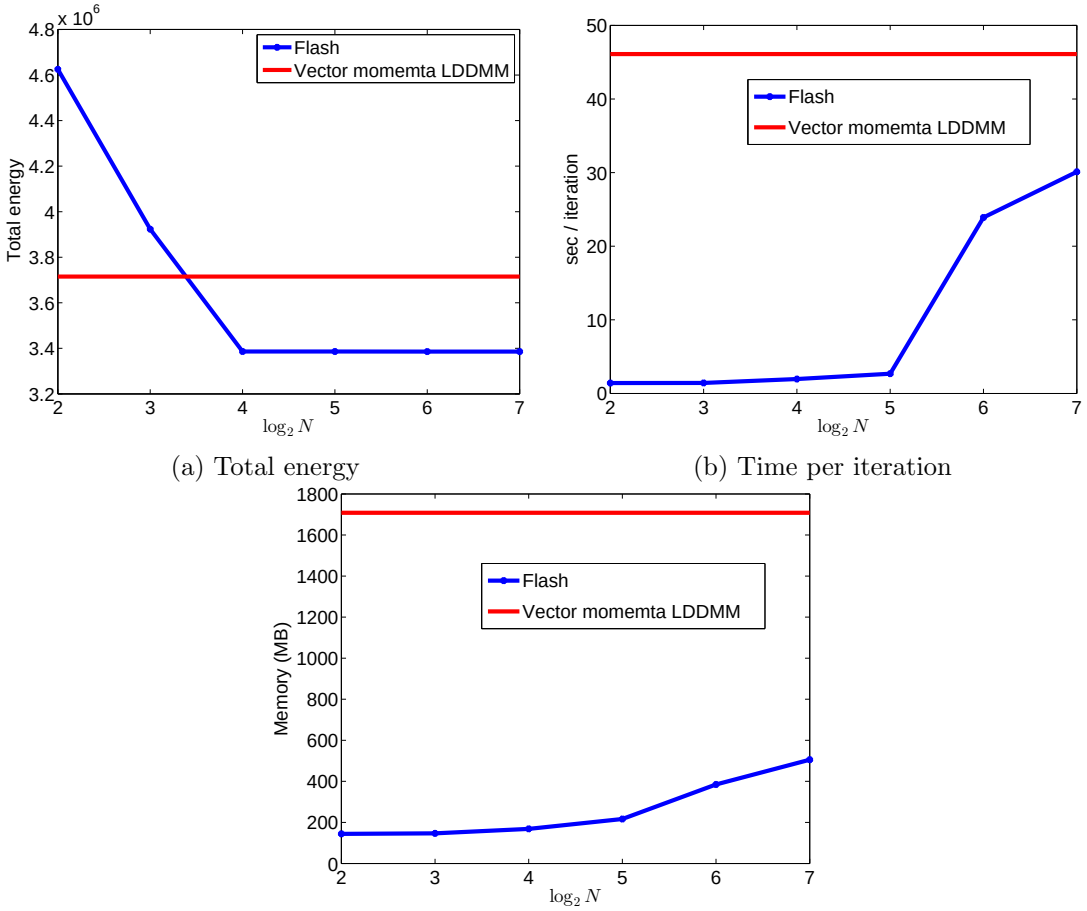


Figure 6.4: Left to right: source image, target images, deformed source image by vector momenta LDDMM, FLASH with $N = 32$. Top to bottom: sine waves with different frequencies $f = 4, 8, 16$.



(a) Total energy

(b) Time per iteration

(c) Memory requirement

Figure 6.5: Comparison between our model FLASH at different scales of truncated dimension and vector momenta LDDMM for (a) total energy, (b) time consumption, and (c) memory requirement.

energy than vector momenta LDDMM for truncated dimension $N = 16$ and higher. Note that increasing the dimension beyond $N = 16$ does not improve the image registration energy, indicating that $N = 16$ is sufficient to capture the transformations between images. We emphasize that we used the same full-dimensional registration energy from (2.15) for all runs so that they would be comparable. In addition, our model FLASH arrives at the optimal solution for $N = 16$ in $1.96s$ per iteration, and 168.4 MB memory. A full dimension of FLASH, $N = 128$, costs 30.1 per iteration and 505.35 MB memory. In comparison, vector momenta LDDMM requires around $46s$ per iteration and 1708.1 MB memory.

Table 6.3 and Table 6.4 give an exact run-time for each step that FLASH ($N = 16$) and vector momenta LDDMM take to compute the gradient term. This provides a practical time cost for both methods that considers the constant factors implied in the big-O complexity notation, as introduced in Section 6.4. It shows that our model FLASH effectively breaks

Table 6.3: Comparison with vector momenta LDDMM for exact run-time on $128\times128\times128$ images with $N = 16$.

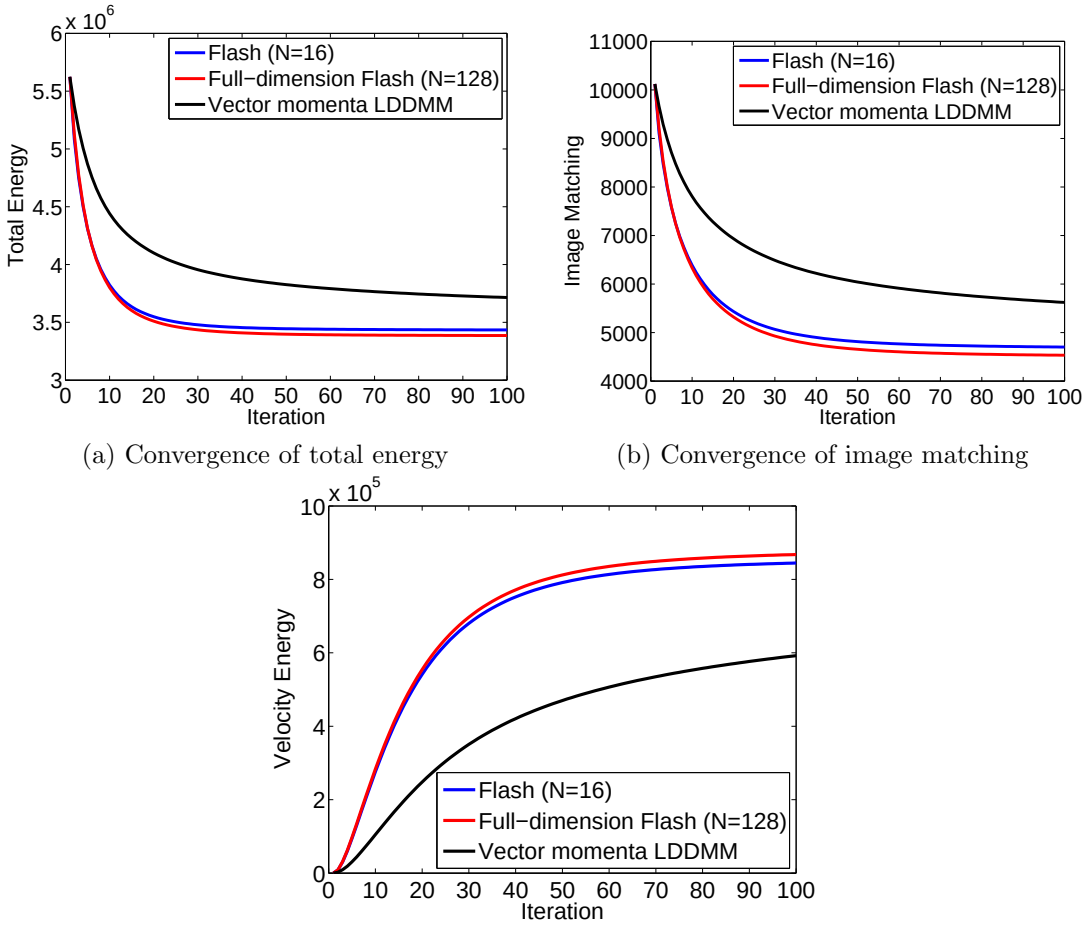
	Run-time	
	FLASH	LDDMM
FWD	$0.1s$	$9.25s$
Compute ϕ	$0s$	$1.5s$
Compute ϕ^{-1}	$1.3s$	$19.0s$
Initial conditions for BWD	$0.16s$	$0.056s$
BWD	$0.40s$	$16s$
Total time	$1.96s$	$45.806s$

Table 6.4: Comparison with vector momenta LDDMM for exact run-time on $256\times256\times256$ images with $N = 16$.

	Run-time	
	FLASH	LDDMM
FWD	$0.12s$	$45s$
Compute ϕ	$0s$	$15.0s$
Compute ϕ^{-1}	$17.64s$	$210.0s$
Initial conditions for BWD	$1.0s$	$0.49s$
BWD	$0.40s$	$140s$
Total time	$19.16s$	$410.49s$

the bottlenecks of vector momenta LDDMM in these three most expensive steps: forward geodesic shooting, inverse deformation field computing, and backward integration. The newly defined low-dimensional Lie algebra moves us to a space that is almost computation-free. The total time speed-up we gain from FLASH at each gradient descent iteration is approximately 23 times faster than vector momenta LDDMM.

We next compared the convergence of FLASH with vector momenta LDDMM. Figure 6.6 demonstrates the convergence graph of total energy, image matching, and velocity energy by FLASH using truncated dimension $N = 16$, full dimension $N = 128$, and vector momenta LDDMM. It shows that FLASH converges not only faster than vector momenta LDDMM, but also to a better optimal solution with lower function energy than in (2.16). The performance of FLASH with $N = 16$ is very close to a full dimension $N = 128$, which means our model does not lose information in a low bandlimited space. Another important fact is that the reduced adjoint Jacobi fields of FLASH completely separate the velocity fields and diffeomorphisms, resulting in the significant advantages of (1) integrating adjoint equations backward in time along geodesics more efficiently in a low-dimensional space, (2) gaining fast convergence.



(a) Convergence of total energy

(b) Convergence of image matching

(c) Convergence of velocity energy

Figure 6.6: Comparison between our model FLASH with $N = 16$ truncated dimension, $N = 128$ full dimension and vector momenta LDDMM for (a) convergence of total energy, (b) convergence of image matching, and (c) convergence of velocity energy.

6.5.3 Atlas Building

We also used FLASH to build an atlas from a set of 3D brain MRIs from the OASIS database, consisting of 60 healthy subjects between the ages of 60 to 95. We initialized the template I as the average of image intensities and set the truncated dimension as $N = 16$, which was shown to be optimal in the previous section. We used a message passing interface (MPI) parallel programming implementation for both our model and vector momenta LDDMM and scattered each image onto an individual processor. With 100 iterations for gradient descent, FLASH builds the atlas in 6.3 minutes, whereas the vector momenta LDDMM in [57] requires 2 hours.

The left side of Figure 6.7 shows the axial and coronal slices from 8 of the selected 3D MRI datasets. The right side shows the atlas image estimated by FLASH, followed by the

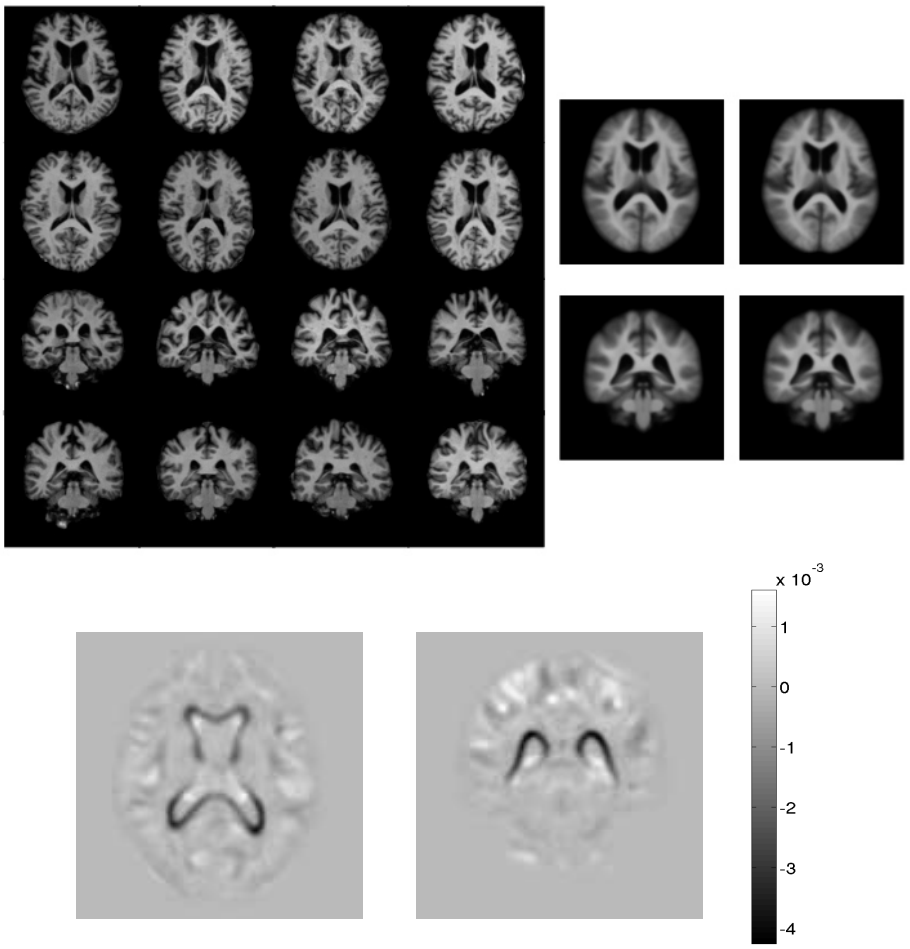


Figure 6.7: Top left: axial and coronal slices from 8 of the input 3D MRIs. Middle to right: atlas estimated by our model with truncated dimension $N = 16$ and vector momenta LDDMM. Bottom: axial and coronal view of atlas intensity difference.

atlas estimated by vector momenta LDDMM. We see from the difference image between the two atlas results that FLASH generated a very similar atlas to the vector momenta LDDMM, but at a fraction of the time and memory cost.

6.6 Conclusion and Future Work

We presented a fast geodesic shooting algorithm, FLASH, for diffeomorphic image registration. Our method is the first to introduce a definition of low-dimensional Lie algebras that provides a discrete representation of the tangent space of diffeomorphisms in a bandlimited velocity space. Another key contribution of this low-dimensional Lie algebra is that we can compute the geodesic evolution equations, as well as the adjoint Jacobi field equations required for gradient descent methods, entirely in a low-dimensional vector space.

This gives us a dramatically fast diffeomorphic image registration algorithm without loss of accuracy. This work can be used for further statistical analysis (for example, anatomical shape variation) since we preserved the distance metric on the space of diffeomorphisms.

There are several interesting directions for future work based on FLASH. First of all, in this chapter, we prespecify the parameters of noise variance, regularization, and the truncated dimension. Since different parameters could lead to a different solution for the diffeomorphic transformation, we can develop a Bayesian formulation much like [58] to estimate all these parameters automatically from the data. This approach would require more expensive computations to treat diffeomorphisms as latent variables and integrate them out from the target distribution, but our algorithm has the ability to make inference via Monte Carlo sampling of the hidden diffeomorphisms more feasible.

Secondly, discovering a connection between this low-dimensional Lie algebra and the diffeomorphism Lie group could be another possible way for further speed up. We have analyzed the computational complexity of our algorithm at each step in Table 6.1 and observed that converting from a low-dimensional Lie algebra to a full dense grid of diffeomorphisms dominates the computational cost. Therefore, making connections between the diffeomorphisms group and our defined discrete Lie algebra in a low-dimensional bandlimited velocity space may ultimately break through the complexity ceiling to the next stage.

In addition, this work also paves the way for efficient computations in large statistical studies using LDDMM. Other speed-up strategies, for instance, a second-order Gauss-Newton step, similar to the one proposed by Ashburner and Friston [6], or a multiresolution optimization scheme, could easily be added on top of our algorithm for further speed improvement.

Finally, we note that the code for FLASH is available as a free library online: (https://bitbucket.org/FlashC/flashc).

CHAPTER 7

DISCUSSION AND FUTURE WORK

This chapter first summarizes the contributions of this dissertation introduced in Chapter 1, followed by a discussion on the possibilities of future work.

7.1 Summary of Contributions

This section reviews the dissertation and claims presented in Chapter 3 to Chapter 6. Each contribution is revisited with a summary of how it was fulfilled.

(1) A Bayesian model of diffeomorphic atlas building has been proposed for the first time to estimate the regularity parameter of transformations. Sampling on the manifold of diffeomorphisms was developed for approximating the expectation step in the inference. This model was then extended to mixture modeling of images from multimode distributions.

A Bayesian diffeomorphic atlas building was introduced in Chapter 3. The major contribution of this work is estimating an atlas simultaneously with the parameters that regularize the smoothness of the diffeomorphic transformations. In other words, the regularity parameter should be selected automatically from the dataset directly rather than manually set. Unlike the previous methods using a mode approximation to alternate between atlas and registration optimizations, a Monte Carlo Expectation Maximization algorithm was developed to marginalize the diffeomorphic image transformations and approximate the expectation step via Hamiltonian Monte Carlo sampling on the manifold of diffeomorphisms.

It was also shown that this stochastic approach successfully solved difficult registration problems involving large deformations where direct geodesic optimization fails. This is because of the fact that the random geodesics of diffeomorphisms generated from HMC sampling prevented the optimization from getting stuck at a local minimum. This single atlas building formulation was later extended to a mixture modeling for estimating multi-atlas that represent subpopulations for images from multimode distributions. In addition, the mixture modeling inference automatically clustered the dataset without knowing the category of each observed image.

*(2) A **generative** Bayesian model of principal geodesic analysis has been developed for automatically selecting the intrinsic dimensions of manifold data. This method provides a compact representation of the data variability, as well as a better geometric fit to the data in a nonlinear space.*

A probabilistic model of principal geodesic analysis on finite-dimensional manifolds was introduced in Chapter 4. This model formed the basis for a Bayesian model of PGA in which the intrinsic dimension of the data can be selected automatically. The model is applicable to generic manifolds. A *Riemannian normal distribution* was carefully defined by the geodesic distance metric. However, due to the lack of an explicit closed-form solution for marginalizing the latent variables on manifolds, we again developed a Monte Carlo Expectation Maximization algorithm where the expectation was approximated by Hamiltonian Monte Carlo sampling of the latent variables. Our Bayesian principal geodesic analysis model selected low-dimensional factors as maximum likelihood. This EM algorithm is computationally efficient in selecting low-dimensional principal factors and avoided having to calculate the sample covariance matrix on manifolds.

(3) The Bayesian principal geodesic analysis was then developed for infinite-dimensional diffeomorphisms. It was shown that reparameterizing the high-dimensional diffeomorphisms in a low-dimensional space captures better intrinsic shape variability of brain MRIs.

The Bayesian principal geodesic analysis of infinite-dimensional diffeomorphisms was presented in Chapter 5. This model provided a probabilistic framework for factor analysis in the space of infinite-dimensional diffeomorphisms and estimated the low-dimensional latent space of diffeomorphic shape variability in a population of images. It was shown that the sparsity prior on the factor matrix resulted in automatic selection of the number of relevant dimensions by pressing irrelevant principal geodesics to zero. The compact representation of diffeomorphisms from our model was able to reconstruct unobserved testing images with lower error than both linear principal component analysis in the image space and tangent space principal component analysis in the diffeomorphism space.

(4) A fast geodesic shooting algorithm for diffeomorphic image registration has been proposed to make the Bayesian inference of diffeomorphisms computationally tractable. In contrast to previous approaches, diffeomorphisms are reparameterized in a discrete low-dimensional bandlimited space rather than a continuous space.

FLASH, a novel definition of low-dimensional Lie algebras in the space of bandlimited velocity fields, was described in Chapter 6. This algorithm was the first to present a real discrete low-dimensional Lie algebra structure that can approximate infinite-dimensional

diffeomorphisms. We computed all the geodesic evolution equations and adjoint Jacobi field equations needed for gradient descent methods entirely in these low-dimensional Lie algebras. In addition, it was shown that a reduced version of adjoint Jacobi field equations used in FLASH gives a faster convergence. We not only sped up the current geodesic shooting algorithm dramatically, but also required much less memory than state-of-the-art methods. The experimental results showed that FLASH was not only faster than state-of-the-art LDDMM algorithms, but also converged to better solutions, i.e., lower values of the registration objective function. This fast geodesic shooting method effectively reduced the high computational cost of computing the gradient term while sampling the diffeomorphisms, and thus made the the Bayesian inference more feasible in time. Hromatka et al. [105], [106] successfully used FLASH in their Bayesian inference for multisite atlases building.

Next, we revisit the thesis statement in Chapter 1.

Thesis: *A **generative** Bayesian approach to analyze the data variability in nonlinear spaces provides parameter estimation and automatic dimensionality reduction for manifold data, such as diffeomorphisms. A Bayesian model of atlas building can for the first time estimate the parameter that regularizes the smoothness of diffeomorphisms. In population studies, (1) reparametrizing diffeomorphisms in a low-dimensional space with appropriate regularity parameters captures better intrinsic shape variability and (2) having a discrete low-dimensional representation of diffeomorphisms makes model inference with Markov Chain Monte Carlo more feasible.*

We showed that developing a Bayesian framework of data variability analysis on manifolds is a natural way to estimate the model parameters and select the number of intrinsic dimensions automatically from the data. Since we had a Bayesian model of atlas building, we were the first to estimate the regularity parameter simultaneously with all other parameters. We then showed that learning an effective and compact representation from the high-dimensional data themselves give a better description of variability. Finally, to make the Bayesian inference with Markov Chain Monte Carlo sampling of diffeomorphisms computationally more tractable, we defined a low-dimensional representation of discretized diffeomorphisms that speeds up diffeomorphic image registration while cutting down its memory requirement.

7.2 Future Work

This section proposes several possibilities to extend the current work for future research. Most of the content was alluded to in the conclusions of Chapters $3, 4, 5$, and 6. We briefly

review the topics here as two parts: Section 7.2.1 describes open theoretical problems about the proposed Bayesian framework, and Sections 7.2.2 describes several extensions, future work of the statistical shape analysis, and other application areas outside of statistical shape analysis and brain MRIs that may benefit from the Bayesian framework developed in this dissertation.

7.2.1 Open Theoretical Problems

There are two major questions regarding our Bayesian framework in this dissertation that remain to be studied.

Does the Hamiltonian Monte Carlo sampling of high-dimensional diffeomorphisms converge? Many approaches are designed to examine the convergence of HMC sampling. Unfortunately, there is no clear way to tell when the Markov chain has converged to its stationary distribution, a.k.a. the target distribution, especially for such high-dimensional data as diffeomorphisms. We can never be sure whether the sampling actually converged. However, we did several tests to check if the chain appears to be converged. We checked the convergence of all estimated parameters by plotting the iteration number against the value of parameters and made sure they did not have bad mixing from a visual aspect. We also ran different chains independently to see whether the expectation value of the target distribution converges to a similar place. More specifically, we checked whether the variance of these expectation values was in a reasonable range.

What is the Lie group associated with our low-dimensional Lie algebra? According to *Lie's third theorem*, there must exist a Lie group associated with this low-dimensional Lie algebra. However, identifying the Lie group still remains an open question. Also, discovering a connection between this discrete Lie algebra and the diffeomorphism Lie group in a low-dimensional space could be another possible way for further speed up. We have analyzed the computational complexity of our algorithm at each step in Table 6.1 and observed that converting from a low-dimensional Lie algebra to a full dense grid of diffeomorphisms back and forth dominates the entire computational cost and memory consumption. Therefore, making this conversion procedure in a low-dimensional space will gain more speed up on top of FLASH.

7.2.2 Related Future Work and Other Applications

The noise model in this dissertation is based on the common assumption of i.i.d. Gaussian at each image voxel. This can be designed differently according to different situations, for instance, a spatially dependent model for correlated noisy data. Similarly, while we chose

a spatially invariant Laplacian for the metric operator L, other metrics such as Gaussian and a wavelet kernel that allow for local regularization properties also fit in our framework. In other words, our model is applicable to spatially varying registration problems.

There are two major aspects of future work. The first is to build upon the Bayesian principal geodesic analysis model for infinite-dimensional diffeomorphisms:

(1) The regularization parameter for the smoothness term of diffeomorphisms was pre-computed using a simple atlas building model in Chapter 3. The reason we did not estimate this parameter simultaneously with all other parameters was the high computational cost of Bayesian inference. Since we developed a fast geodesic shooting algorithm to approximate diffeomorphisms in a low-dimensional space in Chapter 6, we would estimate the regularity parameter jointly with atlas, principal modes, and noise variance. In addition, it would be more natural to use different regularity parameters for different clusters in our mixture model of multiatlas building discussed in Chapter 3, which means each regularity parameter is estimated from its corresponding cluster. The way to compute this parameter is similar to the approach outlined in Chapter 3, but by simply using the data points generated from the same distribution instead of the entire dataset.

(2) Much like the Bayesian principal component analysis model [5] for Euclidean data in linear spaces, we did not enforce the orthogonality of principal modes. This means the estimated principal vectors are not necessarily independent from each other. Despite we considered the orthogonal constraints in Chapter 4 via optimization in the Stiefel manifold of orthonormal frames for finite-dimensional manifold data, the challenge here is that the high-dimensionality of velocity fields makes this a difficult problem to compute directly. However, thanks to our novel definition of low-dimensional Lie algebras introduced in Chapter 6, it hugely reduces the dimensionality in a bandlimited space, which makes the orthogonality constraints much easier to solve.

The other interesting directions for future work are based on the fast geodesic shooting algorithm FLASH, which was introduced in Chapter 6. Instead of learning the parameters of truncated dimension, regularization, and noise variance from the data, we prespecified all of them in FLASH. Ideally, we would automatically estimate the model parameters by developing a Bayesian formulation. Although the Bayesian inference of marginalizing the diffeomorphisms by Monte Carlo sampling requires expensive computations, the developed algorithm FLASH makes this procedure feasible in time. In addition, FLASH paves the way for more efficient computations in large statistical studies using *large deformation diffeomorphic metric mapping*. We can easily add other speed-up strategies, for example, a

second-order Gauss-Newton step, similar to the one proposed in Ashburner and Friston [6], or a multiresolution optimization scheme, on top of FLASH for further speed improvement. We compared the image similarity term in FLASH with the full-dimensional vector momenta LDDMM [57]. However, the accuracy of this diffeomorphic image registration method needs to be more thoroughly evaluated by various methods, for instance, measuring how well the transformed source and target region of labels overlapped [107]. As shown in Chapter 6, FLASH convergences to a lower functional energy than vector momenta LDDMM. It is expected that FLASH will again demonstrate better accuracy, matching the images with a smoother transformation.

The Bayesian framework presented in this dissertation can be very useful for further statistical analysis in the guidance of image segmentation. The statistic models provide shape priors of geometric variability as preknowledge for segmentation tasks. The goal of segmentation is to separate objects of interest in images. In addition to image analysis, this is a fundamental task in many other fields, for instance, computer vision, visualization, etc. The resulting knowledge-based segmentation by considering the shape prior of geometric variability has two major advantages: (1) the shape prior provides more information on ambiguous, missing, or misleading areas; (2) it helps to segment target objects that are corrupted by noise, sampling artifacts, or occlusion.

APPENDIX A

DERIVATIONS FOR BAYESIAN PRINCIPAL GEODESIC ANALYSIS MODEL

Deriving Expectation. The complete expectation function is

$$\begin{aligned} Q(W, x^k, \theta \,|\, \hat{\theta}, \hat{W}) &= E_{\tau \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}} \left[\sum_{k=1}^{N} \log p(W \,|\, J^k; \theta) \right] \\ &\propto -\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left\| I \circ (\phi^k)^{-1} - J^k \right\|_{L^2}^2 - \frac{MN}{2} \log \sigma \\ &\quad - \frac{1}{2} \sum_{k=1}^{N} \left\| W x^k \right\|_V^2 - \sum_{i=1}^{q} \frac{\|W_i\|_V^2}{2} E[\tau_i^{-1} \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}]. \end{aligned} \tag{A.1}$$

Since the sum of log likelihood and the prior on x^k do not depend on τ, we reduce the E-step to compute the conditional expectation $E[\tau_i^{-1} \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}]$. Observe that $p(\tau_i \,|\, J^k; \theta, W, x^k) = p(\tau_i \,|\, W, x^k)$, thus

$$p(\tau_i \,|\, J^k; \theta, W, x^k) = \frac{p(W \,|\, \tau_i, x^k) p(\tau_i)}{\int p(W \,|\, \tau_i, x^k) p(\tau_i) d\tau_i}.$$

The conditional expectation is computed by

$$\begin{aligned} E[\tau_i^{-1} \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}] &= \int \frac{1}{\tau_i} p(\tau_i \,|\, J^k; \hat{\theta}, \hat{W}, \hat{x^k}) d\tau_i, \\ &= \frac{\int \frac{1}{\tau_i} N(\hat{W} \,|\, 0, \tau_i) \frac{1}{\tau_i} d\tau_i}{\int N(\hat{W} \,|\, 0, \tau_i) \frac{1}{\tau_i} d\tau_i}, \\ &= \frac{1}{\|\hat{W}_i\|_V^2}, \end{aligned} \tag{A.2}$$

We obtain the Q function by plugging (A.2) into (A.1).

Deriving Derivatives. Now we compute the variation of $\tilde{Q}$ w.r.t. time-dependent variables I^k, m^k, v^k. Note that the following equations are equivalent for the geodesic paths of each of the subjects, so for notation simplicity, we drop the subject index k momentarily. The derivations are

$$
\begin{aligned}
\partial_I \tilde{Q} &= \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{I}, (\dot{I} + \epsilon \delta \dot{I}) + \nabla (I + \epsilon \delta I) \cdot v \rangle_{L^2} \\
&\quad + \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} \\
&= \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} + \int_0^1 \langle \hat{I}, \delta \dot{I} + \nabla \delta I \cdot v \rangle_{L^2} \\
&= \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} + \langle \hat{I}, \delta I \rangle_{L^2} \Big|_{t=0}^{t=1} - \int_0^1 \langle \dot{\hat{I}}, \delta I \rangle_{L^2} \\
&\quad + \int_0^1 \langle \hat{I}, \nabla \delta I \cdot v \rangle_{L^2} \\
&= \frac{1}{\sigma^2} \langle \delta I_1, I_1 - J \rangle_{L^2} + \langle \hat{I}_1, \delta I_1 \rangle_{L^2} - \langle \hat{I}_0, \delta I_0 \rangle_{L^2} \\
&\quad - \int_0^1 \langle \dot{\hat{I}}, \delta I \rangle_{L^2} - \int_0^1 \langle \nabla \cdot (\hat{I} v), \delta I \rangle_{L^2},
\end{aligned}
$$

$$
\begin{aligned}
\partial_v \tilde{Q} &= \langle Lv, \delta v \rangle_{L^2} + \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{v}, \delta \dot{v} + K \mathrm{ad}^*_{v+\epsilon\delta v} m \rangle_{L^2} \\
&\quad + \langle \hat{I}, \dot{I} + \nabla I \cdot (v + \epsilon \delta v) \rangle_{L^2} \\
&= \langle Lv, \delta v \rangle_{L^2} + \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{v}, \delta \dot{v} \rangle_{L^2} + \langle \mathrm{ad}_{v+\epsilon\delta v} K \hat{v}, m \rangle_{L^2} \\
&\quad + \langle \hat{I}, \dot{I} + \nabla I \cdot (v + \epsilon \delta v) \rangle_{L^2} \\
&= \langle Lv, \delta v \rangle_{L^2} + \langle \hat{v}, \delta v \rangle_{L^2} \Big|_{t=0}^{t=1} - \int_0^1 \langle \dot{\hat{v}}, \delta v \rangle_{L^2} \\
&\quad + \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle -\mathrm{ad}_{K\hat{v}} (v + \epsilon \delta v), m \rangle_{L^2} \\
&\quad + \langle \hat{I}, \dot{I} + \nabla I \cdot (v + \epsilon \delta v) \rangle_{L^2} \\
&= \langle Lv, \delta v \rangle_{L^2} + \langle \hat{v}_1, \delta v_1 \rangle_{L^2} - \langle \hat{v}_0, \delta v \rangle_{L^2} \\
&\quad + \int_0^1 \langle -\mathrm{ad}_{K\hat{v}} \delta v, m \rangle_{L^2} + \langle \hat{I}, \nabla I \cdot \delta v \rangle_{L^2} - \langle \dot{\hat{v}}, \delta v \rangle_{L^2} \\
&= \langle Lv, \delta v \rangle_{L^2} + \langle \hat{v}_1, \delta v_1 \rangle_{L^2} - \langle \hat{v}_0, \delta v \rangle_{L^2} \\
&\quad + \int_0^1 \langle -\mathrm{ad}^*_{K\hat{v}} m, \delta v \rangle_{L^2} + \langle \hat{I} \nabla I, \delta v \rangle_{L^2} - \langle \dot{\hat{v}}, \delta v \rangle_{L^2},
\end{aligned}
$$

$$
\begin{aligned}
\partial_m \tilde{Q} &= \frac{\partial}{\partial \epsilon}\Big|_{\epsilon=0} \int_0^1 \langle \hat{v}, K \mathrm{ad}^*_v (m + \epsilon \delta m) \rangle_{L^2} + \langle \hat{m}, (m + \epsilon \delta m) \rangle_{L^2} \\
&= \int_0^1 \langle \mathrm{ad}_v K \hat{v}, \delta m \rangle_{L^2} + \langle \hat{m}, \delta m \rangle_{L^2},
\end{aligned}
$$

here $\nabla\cdot$ is the divergence operator. Since we have $\delta I_0 = 0$, $\delta v = 0$, the optimality conditions for I, v are given by the following time-dependent system of ODEs, termed the *adjoint equations*:

$$\left.\begin{aligned} -\dot{\hat{I}} - \nabla\cdot(\hat{I}v) &= 0, \\ -\mathrm{ad}^*_{K\hat{v}}m + \hat{I}\nabla I - \dot{\hat{v}} &= 0, \\ \mathrm{ad}_v K\hat{v} + \hat{m} &= 0, \end{aligned}\right\} \tag{A.3}$$

subject to initial conditions

$$\hat{v}_1 = 0, \qquad \hat{I}_1 = \frac{1}{\sigma^2}(I_1 - J).$$

Finally, after integrating these adjoint equations backwards in time to $t = 0$, the gradient of $\tilde{Q}$ with respect to the kth initial velocity is

$$\nabla_{v^k}\tilde{Q} = v^k - K\hat{v}^k.$$

Deriving Closed-form Solution for θ. Notice that the gradient term of (5.6) w.r.t. $\theta = \{I, \sigma\}$ relates only to the image matching term

$$\tilde{Q}(I,\sigma) = -\frac{1}{2\sigma^2}\sum_{k=1}^{N}\left\|I\circ(\phi^k)^{-1} - J^k\right\|^2_{L^2} - \frac{MN}{2}\log\sigma. \tag{A.4}$$

We have the gradient of (A.4) as

$$\partial_\sigma\tilde{Q}(I,\sigma) = \frac{1}{\sigma^3}\sum_{k=1}^{N}\left\|I\circ(\phi^k)^{-1} - J^k\right\|^2_{L^2} - \frac{MN}{2\sigma}. \tag{A.5}$$

Setting (A.5) to zero, we then get the closed-form formulation to update σ^2 by

$$\sigma^2 = \frac{1}{MN}\sum_{k=1}^{N}\left\|I\circ(\phi^k)^{-1} - J^k\right\|^2_{L^2}.$$

Next, we compute the gradient of (5.6) w.r.t. the atlas I by changing variables $y = \left(\phi^k\right)^{-1}(x)$, such that

$$x = \phi^k(y), \quad dx = |D\phi^k(y)|dy.$$

After dropping the normalizing constant of σ which is irrelevant to I, we expand and rewrite equation (A.4) as

$$\tilde{Q}(I) = \sum_{k=1}^{N}\int_\Omega \langle I(y) - J^k\circ\phi^k(y), I(y) - J^k\circ\phi^k(y)\rangle_{L^2}\,|D\phi^k(y)|dy.$$

This gives the derivative w.r.t I by

$$\partial_I \tilde{Q} = \sum_{k=1}^{N} (I - J^k \circ \phi^k) |D\phi^k| \tag{A.6}$$

Equating (A.6) to zero at optimal, we have

$$I = \frac{\sum_{k=1}^{N} J^k \circ \phi^k |D\phi^k|}{\sum_{k=1}^{N} |D\phi^k|}.$$

APPENDIX B

SUPPLEMENT TO LOW-DIMENSIONAL LIE ALGEBRAS

Properties of truncated convolution. We compute the truncated convolution in a bandlimited space by padding a sufficient number of zeros, $N_i - 1$, at the end of each dimension, and then truncating the signal back to its original space. Note that (1) the convolution is modulo-N_i circular convolution, (2) the zero frequency component is shifted to the center of the domain.

We first introduce the kth element of a truncated convolution of any tangent vector field $\tilde{v}, \tilde{w} \in \tilde{V}$ as

$$[\tilde{v} * \tilde{w}]_k = \sum_{0 \leqslant l < 2N} \tilde{v}_{k-l} \tilde{w}_l, \tag{B.1}$$

where $l = (l_1, \ldots, l_d) \in \mathbb{Z}^d$ are the padded frequency indexes with $l_i \in \{0, \ldots, 2N_i - 1\}$ along the ith axis. Let $M_i = \lfloor \frac{N_i - 1}{2} \rfloor$, then $k = (k_1, \ldots, k_d) \in \mathbb{Z}^d$ denotes truncated frequency indexes with $k_i \in \{M_i, \ldots, M_i + N_i - 1\}$. Notice here the $k - l$ represents $k - l \,(\mod N)$ with cyclic boundary conditions, and the vector field $\tilde{v}$ can also be replaced with a matrix field.

We then prove the commutativity and associativity of this truncated convolution. For any $\tilde{u}, \tilde{v}, \tilde{w} \in \tilde{V}$,

- Commutativity: $\tilde{u} * \tilde{v} = \tilde{v} * \tilde{u}$

  Proof: The kth element of $\tilde{u} * \tilde{v}$ is

  $$[\tilde{u} * \tilde{v}]_k = \sum_{0 \leqslant l < 2N} \tilde{u}_{k-l} \tilde{v}_l.$$

  By changing the coordinates using $k - l = k'$ where $k \in [M, M + N)$ and $k' \in [0, 2N)$ due to the cyclic condition, we rewrite the equation above as

  $$[\tilde{u} * \tilde{v}]_k = \sum_{0 \leqslant k' < 2N} \tilde{u}_{k'} \tilde{v}_{k-k'} = [\tilde{v} * \tilde{u}]_k.$$

- Associativity: $(\tilde{u} * \tilde{v}) * \tilde{w} = \tilde{u} * (\tilde{v} * \tilde{w})$

  Proof: The kth element of $(\tilde{u} * \tilde{v}) * \tilde{w}$ is

  $$[(\tilde{u} * \tilde{v}) * \tilde{w}]_k = \sum_{0 \leqslant l' < 2N} \sum_{0 \leqslant l < 2N} \tilde{u}_{k-l'-l} \tilde{v}_l \tilde{w}_{l'},$$

  where $k \in [M, M+N)$. We then rewrite the equation above by changing the coordinates using $k - l' - l = a$, where $a \in [0, 2N)$, as

  $$[(\tilde{u} * \tilde{v}) * \tilde{w}]_k = \sum_{0 \leqslant l' < 2N} \sum_{0 \leqslant a < 2N} \tilde{u}_a \tilde{v}_{k-a-l'} \tilde{w}_{l'} = [\tilde{u} * (\tilde{v} * \tilde{w})]_k,$$

Deriving ad* **operator.** Before deriving the ad* operator, we write the pairing between the truncated convolution (B.1) and a momentum vector field $\tilde{m} \in \tilde{V}^*$ as

$$\begin{aligned}(\tilde{m}, \tilde{v} * \tilde{w}) &= \sum_{M \leqslant k < M+N} (\tilde{m}_{k-M}, \sum_{0 \leqslant l < 2N} \tilde{v}_{k-l} \tilde{w}_l) \\ &= \sum_{M \leqslant k < M+N} \sum_{0 \leqslant l < 2N} (\tilde{m}_{k-M}, \tilde{v}_{k-l} \tilde{w}_l)\end{aligned}$$

We are now ready to derive ad* operator from its definition

$$(\mathrm{ad}^*_{\tilde{v}} \tilde{m}, \tilde{w}) = (\tilde{m}, \mathrm{ad}_{\tilde{v}} \tilde{w}).$$

Plugging the Lie bracket introduced in (6.4), we have

$$\begin{aligned}(\mathrm{ad}^*_{\tilde{v}} \tilde{m}, \tilde{w}) &= \Big(\tilde{m}, (\tilde{D}\tilde{v}) * \tilde{w} - (\tilde{D}\tilde{w}) * \tilde{v}\Big) \\ &= \Big(\tilde{m}, (\tilde{D}\tilde{v}) * \tilde{w}\Big) - \Big(\tilde{m}, (\tilde{D}\tilde{w}) * \tilde{v}\Big) \\ &= \sum_{M \leqslant k < M+N} \sum_{0 \leqslant l < 2N} (\tilde{m}_{k-M}, \tilde{v}_{k-l} \tilde{\eta}^T_{k-l} \tilde{w}_l) - (\tilde{m}_{k-M}, \tilde{w}_l \tilde{\eta}^T_l \tilde{v}_{k-l})\end{aligned}$$

To separate $\tilde{w}$, we change coordinates by defining $k - M = k'$, $l - M = l'$ where $k' \in [0, N-1]$ and $l' \in [-M, 2N-1-M]$. For the purpose of notation simplicity, we drop the range of variables in the following equations as

$$\begin{aligned}(\mathrm{ad}^*_{\tilde{v}} \tilde{m}, \tilde{w}) &= \sum_{k'} \sum_{l'} (\tilde{m}_{k'}, \tilde{v}_{k'-l'} \tilde{\eta}^T_{k'-l'} \tilde{w}_{l'+M}) - (\tilde{m}_{k'}, \tilde{w}_{l'+M} \tilde{\eta}^T_{l'+M} \tilde{v}_{k'-l'}) \\ &= \sum_{k'} \sum_{l'} \big((\tilde{v}_{k'-l'} \tilde{\eta}^T_{k'-l'})^{T*} \tilde{m}_{k'}, \tilde{w}_{l'+M}\big) - (\tilde{v}^*_{k'-l'} \tilde{m}^T_{k'}, \tilde{w}_{l'+M} \tilde{\eta}^T_{l'+M}) \\ &= \sum_{k'} \sum_{l'} \big((\tilde{v}_{k'-l'} \tilde{\eta}^T_{k'-l'})^{T*} \tilde{m}_{k'}, \tilde{w}_{l'+M}\big) - (\tilde{v}^*_{k'-l'} \tilde{m}^T_{k'} \tilde{\eta}^*_{l'+M}, \tilde{w}_{l'+M}) \\ &= \sum_{k'} \sum_{l'} \big((\tilde{v}_{k'-l'} \tilde{\eta}^T_{k'-l'})^{T*} \tilde{m}_{k'}, \tilde{w}_{l'+M}\big) + (\tilde{v}^*_{k'-l'} \tilde{m}^T_{k'} \tilde{\eta}_{l'+M}, \tilde{w}_{l'+M}) \\ &= \Big((\tilde{D}\tilde{v})^T \star \tilde{m}, \tilde{w}\Big) + \Big(\tilde{\Gamma}(\tilde{m} \otimes \tilde{v}), \tilde{w}\Big) \\ &= \Big((\tilde{D}\tilde{v})^T \star \tilde{m} + \tilde{\Gamma}(\tilde{m} \otimes \tilde{v}), \tilde{w}\Big).\end{aligned}$$

REFERENCES

[1] F.-X. Vialard, L. Risser, D. Rueckert, and C. J.Cotter, "Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation," in *Int. J. Comput. Vision*, 2012, pp. 229–241.

[2] S. Allassonnière, Y. Amit, and A. Trouvé, "Toward a coherent statistical framework for dense deformable template estimation," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 69, pp. 3–29, 2007.

[3] M. Vaillant, M. I. Miller, L. Younes, and A. Trouvé, "Statistics on diffeomorphisms via tangent space representations," *NeuroImage*, vol. 23, pp. S161–S169, 2004.

[4] A. Qiu, L. Younes, and M. I. Miller, "Principal component based diffeomorphic surface mapping," *IEEE Trans. Med. Imaging*, vol. 31, no. 2, pp. 302–311, 2012.

[5] C. M. Bishop, "Bayesian PCA," *Adv. Neural Inf. Process Syst.*, pp. 382–388, 1999.

[6] J. Ashburner and K. J. Friston, "Diffeomorphic registration using geodesic shooting and Gauss-Newton optimisation," *NeuroImage*, vol. 55, no. 3, pp. 954–967, 2011.

[7] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," *IEEE Trans. Image Process.*, vol. 5, no. 10, pp. 1435–1447, 1996.

[8] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, "Diffeomorphic demons: Efficient non-parametric image registration," *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.

[9] K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009, vol. 494.

[10] D. G. Kendall, "Shape manifolds, Procrustean metrics, and complex projective spaces," *Bull. London Math. Soc.*, vol. 16, pp. 18–121, 1984.

[11] N. Courty, T. Burger, and P.-F. Marteau, "Geodesic analysis on the gaussian rkhs hypersphere," in *ECML-PKDD*. Springer, 2012, pp. 299–313.

[12] P. T. Fletcher and S. Joshi, "Principal geodesic analysis on symmetric spaces: statistics of diffusion tensors," in *CVAMIA*, 2004.

[13] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, 2008.

[14] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2273–2286, 2011.

[15] W. M. Boothby, *An introduction to differentiable manifolds and Riemannian geometry.* Gulf Professional Publishing, 2003, vol. 120.

[16] M. do Carmo, *Riemannian Geometry.* Birkhäuser, 1992.

[17] J. Cheeger and D. G. Ebin, *Comparison theorems in Riemannian geometry.* Amer. Math. Soc., 2008, vol. 365.

[18] J. J. Duistermaat and J. A. Kolk, *Lie groups.* Springer Science & Business Media, New York, 2012.

[19] B. Hall, *Lie groups, Lie algebras, and representations: an elementary introduction.* Springer, New York, 2015, vol. 222.

[20] B. Francesco, "Invariant affine connections and controllability on lie groups," technical Report for Geometric Mechanics, California Institute of Technology, Tech. Rep., 1995.

[21] J. Hinkle, P. T. Fletcher, and S. Joshi, "Intrinsic polynomials for regression on riemannian manifolds," *J. Math. Imaging. Vis.*, vol. 50, no. 1-2, pp. 32–52, 2014.

[22] I. T. Jolliffe, *Principal Component Analysis.* Springer-Verlag New York, 1986, vol. 487.

[23] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. R. Stat. Soc. Series B Stat. Methodol.*, vol. 61, no. 3, pp. 611–622, 1999.

[24] S. Roweis, "EM algorithms for PCA and SPCA," *Adv. Neural Inf. Process Syst.*, pp. 626–632, 1998.

[25] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Department of Statistics, University of California, Berkeley, Tech. Rep. 608, 2005.

[26] N. D. Lawrence, "Gaussian process latent variable models for visualisation of high dimensional data," *Adv. Neural Inf. Process Syst.*, vol. 16, pp. 329–336, 2004.

[27] P. T. Fletcher, C. Lu, and S. Joshi, "Statistics of shape via principal geodesic analysis on Lie groups," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, 2003, pp. 95–101.

[28] M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," in *Ann. I. H. Poincare*, vol. 10. Presses universitaires de France, 1948, pp. 215–310.

[29] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Comm. Pure Appl. Math.*, vol. 30, no. 5, pp. 509–541, 1977.

[30] W. S. Kendall, "Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence," *P. Lond. Math. Soc.*, vol. 3, no. 61, pp. 371–406, 1990.

[31] X. Pennec, "Probabilities and statistics on Riemannian manifolds: Basic tools for geometric measurements," in *IEEE Workshop on NSIP*, vol. 4. Citeseer, 1999.

[32] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans. Med. Imaging*, vol. 23, no. 8, pp. 995–1005, 2004.

[33] P. T. Fletcher, "Geodesic regression on Riemannian manifolds," in *Med. Image Comput. Comput.-Assisted Intervention Workshop on MFCA*, 2011, pp. 75–86.

[34] ——, "Geodesic regression and the theory of least squares on riemannian manifolds," *Int. J. Comput. Vision*, vol. 105, no. 2, pp. 171–185, 2013.

[35] M. Niethammer, Y. Huang, and F.-X. Viallard, "Geodesic regression for image time-series," in *Med. Image Comput. Comput.-Assisted Intervention*, 2011.

[36] J. Hinkle, P. T. Fletcher, and S. C. Joshi, "Intrinsic polynomials for regression on riemannian manifolds," *J. Math. Imaging. Vis.*, vol. 50, no. 1–2, pp. 32–52, 2014.

[37] B. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi, "Population shape regression from random design data," in *IEEE I. Conf. Comp. Vis.*, 2007.

[38] X. Shi, M. Styner, J. Lieberman, J. Ibrahim, W. Lin, and H. Zhu, "Intrinsic regression models for manifold-valued data," in *Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2009, pp. 192–199.

[39] J. Su, I. L. Dryden, E. Klassen, H. Le, and A. Srivastava, "Fitting smoothing splines to time-indexed, noisy points on nonlinear manifolds," *Image Vis. Comput.*, vol. 30, no. 6, pp. 428–442, 2012.

[40] P. E. Jupp and J. T. Kent, "Fitting smooth paths to spherical data," *Appl. Stat.*, vol. 36, no. 1, pp. 34–46, 1987.

[41] A. Kume, I. L. Dryden, and H. Le, "Shape-space smoothing splines for planar landmark data," *Biometrika*, vol. 94, no. 3, pp. 513–528, 2007.

[42] M. Miller, "Computational anatomy: shape, growth, and atrophy comparison via diffeomorphisms," *NeuroImage*, vol. 23, pp. S19–S33, 2004.

[43] A. Trouvé and F.-X. Vialard, "A second-order model for time-dependent data interpolation: Splines on shape spaces," in *Med. Image Comput. Comput.-Assisted Intervention STIA Workshop*, 2010.

[44] S. Durrleman, X. Pennec, A. Trouvé, G. Gerig, and N. Ayache, "Spatiotemporal atlas estimation for developmental delay detection in longitudinal datasets," in *Med. Image Comput. Comput.-Assisted Intervention*, 2009, pp. 297–304.

[45] L. Noakes, G. Heinzinger, and B. Paden, "Cubic splines on curved spaces," *IMA J. Math. Contr. Inform.*, vol. 6, no. 4, pp. 465–473, 1989.

[46] P. Crouch and F. S. Leite, "The dynamic interpolation problem: on Riemannian manifolds, Lie groups, and symmetric spaces," *J. Dyn. Control Syst.*, vol. 1, no. 2, pp. 177–202, 1995.

[47] S. R. Buss and J. P. Fillmore, "Spherical averages and applications to spherical splines and interpolation," *ACM Trans. Graph.*, vol. 20, no. 2, pp. 95–126, 2001.

[48] V. I. Arnol'd, "Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits," *Ann. Inst. Fourier*, vol. 16, pp. 319–361, 1966.

[49] M. I. Miller, A. Trouvé, and L. Younes, "Geodesic shooting for computational anatomy," *J. Math. Imaging Vis.*, vol. 24, no. 2, pp. 209–228, 2006.

[50] P. Dupuis, U. Grenander, and M. I. Miller, "Variational problems on flows of diffeomorphisms for image matching," *Q. J. Math.*, vol. 56, no. 3, p. 587, 1998.

[51] A. Trouvé, "Diffeomorphisms groups and pattern matching in image analysis," *Int. J. Comput. Vision*, vol. 28, no. 3, pp. 213–221, 1998.

[52] M. Beg, M. Miller, A. Trouvé, and L. Younes, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *Int. J. Comput. Vision*, vol. 61, no. 2, pp. 139–157, 2005.

[53] M. Zhang and P. T. Fletcher, "Probabilistic principal geodesic analysis," in *Adv. Neural Inf. Process Syst.*, 2013, pp. 1178–1186.

[54] ——, "Bayesian principal geodesic analysis in diffeomorphic image registration," in *Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2014, pp. 121–128.

[55] ——, "Bayesian principal geodesic analysis for estimating intrinsic diffeomorphic image variability," *Med. Image Anal.*, vol. 25, no. 1, pp. 37–44, 2015.

[56] L. Younes, F. Arrate, and M. Miller, "Evolutions equations in computational anatomy," *NeuroImage*, vol. 45, no. 1S1, pp. 40–50, 2009.

[57] N. Singh, J. Hinkle, S. Joshi, and P. T. Fletcher, "A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction," in *Proc. Int. Symp. Biomed. Imag.*, April 2013.

[58] M. Zhang, N. Singh, and P. T. Fletcher, "Bayesian estimation of regularization and atlas building in diffeomorphic image registration," in *Inf. Process. Med. Imaging*. Springer, 2013, pp. 37–48.

[59] M. Zhang, H. Shao, and P. T. Fletcher, "A mixture model for automatic diffeomorphic multi-atlas building," in *Med. Image Comput. Comput.-Assisted Intervention Workshop - BAMBI*, 2015.

[60] K. Bhatia, J. Hajnal, B. Puri, A. Edwards, and D. Rueckert, "Consistent groupwise non-rigid registration for atlas construction," in *Proc. Int. Symp. Biomed. Imag.*, 2004.

[61] L. Zöllei, M. Jenksinon, S. Timoner, and W. W. III, "A marginalized map approach and em optimization for pair-wise registration," in *Inf. Process. Med. Imaging*, 2007, pp. 662–667.

[62] S. Joshi, B. Davis, M. Jomier, and G. Gerig, "Unbiased diffeomorphic atlas construction for computational anatomy," *NeuroImage*, vol. 23, Supplement1, pp. 151–160, 2004.

[63] C. Twining, T. Cootes, S. Marsland, V. Petrovic, R. Schestowitz, and C. Taylor, "A unified information-theoretic approach to groupwise non-rigid registration and model building," in *Inf. Process. Med. Imaging*, 2005, pp. 1–14.

[64] F.-X. Vialard, L. Risser, D. Holm, and D. Rueckert, "Diffeomorphic atlas estimation using Kärcher mean and geodesic shooting on volumetric images," in *MIUA*, 2011.

[65] S. Allassonnière and E. Kuhn, "Stochastic algorithm for parameter estimation for dense deformable template mixture model," *ESAIM-PS*, vol. 14, pp. 382–408, 2010.

[66] K. V. Leemput, "Encoding probabilistic brain atlases using Bayesian inference," *IEEE Trans. Med. Imaging*, vol. 28, pp. 822–837, 2009.

[67] J. E. Iglesias, M. R. Sabuncu, K. V. Leemput, and ADNI, "Incorporating parameter uncertainty in Bayesian segmentation models: application to hippocampal subfield volumetry," in *Med. Image Comput. Comput.-Assisted Intervention*, 2012.

[68] P. Risholm, S. Pieper, E. Samset, and W. M. Wells, "Summarizing and visualizing uncertainty in non-rigid registration," in *Med. Image Comput. Comput.-Assisted Intervention*, 2010.

[69] P. Risholm, E. Samset, and W. M. Wells, "Bayesian estimation of deformation and elastic parameters in non-rigid registration," in *WBIR*, 2010.

[70] I. J. A. Simpson, J. A. Schnabel, A. R. Groves, J. L. R. Andersson, and M. W. Woolrich, "Probabilistic inference of regularisation in non-rigid registration," *NeuroImage*, vol. 59, pp. 2438–2451, 2012.

[71] J. Ma, M. I. Miller, A. Trouvé, and L. Younes, "Bayesian template estimation in computational anatomy," *NeuroImage*, vol. 42, pp. 252–261, 2008.

[72] A. Budhiraja, P. Dupuis, and V. Maroulas, "Large deviations for stochastic flows of diffeomorphisms," *Bernoulli*, vol. 16, pp. 234–257, 2010.

[73] S. Duane, A. Kennedy, B. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Phys. Lett. B*, pp. 216–222, 1987.

[74] D. J. Blezek and J. V. Miller, "Atlas stratification," in *Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2006, pp. 712–719.

[75] M. R. Sabuncu, S. K. Balci, M. E. Shenton, and P. Golland, "Image-driven population analysis through mixture modeling," *IEEE Trans. Med. Imaging*, vol. 28, no. 9, pp. 1473–1487, 2009.

[76] X. Tang, K. Oishi, A. V. Faria, A. E. Hillis, M. S. Albert, S. Mori, and M. I. Miller, "Bayesian parameter estimation and segmentation in the multi-atlas random orbit model," *PloS One*, vol. 8, no. 6, p. e65591, 2013.

[77] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy," *Neuroimage*, vol. 46, no. 3, pp. 726–738, 2009.

[78] J. Koikkalainen, J. Lötjönen, L. Thurfjell, D. Rueckert, G. Waldemar, H. Soininen, and ADNI, "Multi-template tensor-based morphometry: application to analysis of Alzheimer's disease," *NeuroImage*, vol. 56, no. 3, pp. 1134–1144, 2011.

[79] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains," *NeuroImage*, vol. 21, no. 4, pp. 1428–1442, 2004.

[80] E. M. Van Rikxoort, I. Isgum, Y. Arzhaeva, M. Staring, S. Klein, M. A. Viergever, J. P. Pluim, and B. van Ginneken, "Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus," *Med. Image Anal.*, vol. 14, no. 1, pp. 39–49, 2010.

[81] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, D. Rueckert, and ADNI, "LEAP: learning embeddings for atlas propagation," *NeuroImage*, vol. 49, no. 2, pp. 1316–1325, 2010.

[82] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, 2007.

[83] P. T. Fletcher and M. Zhang, "Probabilistic geodesic models for regression and dimensionality reduction on riemannian manifolds," in *RCCV*. Springer, 2016, pp. 101–121.

[84] A. Bhattacharya and D. B. Dunson, "Nonparametric bayesian density estimation on manifolds with applications to planar shapes," *Biometrika*, vol. 97, no. 4, pp. 851–865, 2010.

[85] S. Byrne and M. Girolami, "Geodesic Monte Carlo on embedded manifolds," *arXiv preprint arXiv:1301.6064*, 2013.

[86] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[87] S. Said, N. Courty, N. Le Bihan, and S. J. Sangwine, "Exact principal geodesic analysis for data on SO(3)," in *15th Eur. Signal Process. Conf.*, 2007, pp. 1700–1705.

[88] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen, "Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 43–56.

[89] S. Huckemann and H. Ziezold, "Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces," *Adv. in Appl. Probab.*, vol. 38, no. 2, pp. 299–319, 2006.

[90] S. Jung, I. L. Dryden, and J. S. Marron, "Analysis of principal nested spheres," *Biometrika*, vol. 99, no. 3, pp. 551–568, 2012.

[91] X. Pennec, "Intrinsic statistics on Riemannian manifolds: basic tools for geometric measurements," *J. Math. Imaging Vis.*, vol. 25, no. 1, 2006.

[92] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal.*, vol. 20, no. 2, pp. 303–353, 1998.

[93] P. Gori, O. Colliot, Y. Worbe, L. Marrakchi-Kacem, S. Lecomte, C. Poupon, A. Hartmann, N. Ayache, and S. Durrleman, "Bayesian atlas estimation for the variability analysis of shape complexes," in *Med. Image Comput. Comput.-Assisted Intervention*, vol. 8149. Springer, 2013, pp. 267–274.

[94] M. A. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, 2003.

[95] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, 2007.

[96] S. Gerber, T. Tasdizen, P. T. Fletcher, S. Joshi, and R. Whitaker, "Manifold modeling for brain population analysis," *Med. Image Anal.*, vol. 14, no. 5, pp. 643–653, 2010.

[97] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *J. R. Stat. Soc. Series B Stat. Methodol.*, pp. 99–102, 1974.

[98] K. Van Leemput, "Encoding probabilistic brain atlases using Bayesian inference," *IEEE Trans. Med. Imaging*, vol. 28, no. 6, pp. 822–837, 2009.

[99] M. Zhang and P. T. Fletcher, "Finite-dimensional Lie algebras for fast diffeomorphic image registration," in *Inf. Process. Med. Imaging*. Springer, 2015, pp. 249–260.

[100] V. Arsigny, O. Commowick, X. Pennec, and N. Ayache, "A log-Euclidean framework for statistics on diffeomorphisms," in *Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2006, pp. 924–931.

[101] J. Ashburner, "A fast diffeomorphic image registration algorithm," *Neuroimage*, vol. 38, no. 1, pp. 95–113, 2007.

[102] N. Singh, P. T. Fletcher, J. S. Preston, L. Ha, R. King, J. S. Marron, M. Wiener, and S. Joshi, "Multivariate statistical analysis of deformation momenta relating anatomical shape to neuropsychological measures," in *Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2010, pp. 529–537.

[103] S. Durrleman, M. Prastawa, G. Gerig, and S. Joshi, "Optimal data-driven sparse parameterization of diffeomorphisms for population analysis," in *Inf. Process. Med. Imaging*. Springer, 2011, pp. 123–134.

[104] T. Schmah, L. Risser, and F.-X. Vialard, "Diffeomorphic image matching with left-invariant metrics," in *Geometry, Mechanics, and Dynamics*. Springer, 2015, pp. 373–392.

[105] M. Hromatka, M. Zhang, G. M. Fleishman, B. Gutman, N. Jahanshad, P. Thompson, and P. T. Fletcher, "A hierarchical bayesian model for multi-site diffeomorphic image atlases," in *Med. Image Comput. Comput.-Assisted Intervention*. Springer, 2015, pp. 372–379.

[106] Y. Gao, M. Zhang, P. T. Fletcher, and G. Gerig, "Image registration and segmentation in longitudinal MRI using temporal appearance modeling," in *Proc. Int. Symp. Biomed. Imag.* Springer, 2016.

[107] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration," *Neuroimage*, vol. 46, no. 3, pp. 786–802, 2009.